Fall 8-12-2014

A meta-analysis of Type I error rates for detecting differential item functioning with logistic regression and Mantel-Haenszel in Monte Carlo studies

Eva Van De Water

ACCEPTANCE

A META-ANALYSIS OF TYPE I ERROR RATES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITH LOGISTIC REGRESSION AND MANTEL-HAENSZEL IN MONTE CARLO STUDIES, by Eva C. Van De Water was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____	_____
William Curlette, Ph.D. Committee Chair	Chris T. Oshima, Ph.D. Committee Member
_____	_____
Theresa A. Sipe, Ph.D. Committee Member	Meltem Alemdar, Ph.D. Committee Member

_____
Hongli Li, Ph.D.
Committee Member

_____
Date

_____
William L. Curlette, Ph.D.
Chairperson, Department of Educational Policy Studies

_____
Paul A. Alberto, Ph.D.
Dean
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Eva C. Van De Water

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Eva C. Van De Water
1191 Burnt Creek Court
Decatur, Georgia 30033

The director of this dissertation is:

Dr. William L. Curlette
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, Georgia 30303-3083

CURRICULUM VITA

Eva Courtney Van De Water

ADDRESS: 1191 Burnt Creek Court
Decatur, GA 30033

EDUCATION:

Ph. D. 2012	Georgia State University Education Policy Studies
Cert 2005	Gifted Infield 6-12 ESOL
Cert. 2001	T-6
Ed. S. 2001	Georgia College and State University Natural Science Education
M.A.T. 1997	Georgia State University Broad Field Science Education
B.S. 1994	Valdosta State University Biology

PROFESSIONAL EXPERIENCE:

2011-Present	Psychometrician ALTA Language Services, Atlanta, GA
2010- 2011	Environmental Science Teacher Arabia Mountain High School, Lithonia, GA
2008-2010	Physics & Chemistry Teacher Druid Hills High School, Atlanta, GA
2006-2008	Graduate Research Assistant, Graduate Teaching Assistant Georgia State University, Atlanta, GA
2004-2005	Biology Teacher and Department Chair Berkmar Middle School, Lilburn, GA
2002-2004	Physical Science, Geometry & Spanish Teacher Westside High School, Macon, GA
2002	Research Coordinator, Surgical Research Mercer University School of Medicine, Macon, GA
1997-2001	Biology & Chemistry Teacher Westside High School, Macon, GA

PRESENTATIONS AND PUBLICATIONS:

Gowen, S., Furlow, C., Skelton, S., Lingle, J., Olowoye, S., Davis, C., & Van De Water, E. (2007). *Evaluation of Georgia's 21st Century Community Learning Centers: Phase II Report*. Georgia Department of Education.

Lingle, J., Alemdar, M., Gowen, S., Fournillier, J., Skelton, S., Davis, C., Van De Water, E., Olowoye, S. (2008). *School Engagement and Healthy Behaviors: Results from the Evaluation Community-Based After-School Programs*. Paper presentation at the Annual Conference for the American Educational Research Association, New York, NY.

Weerasuriya, A., Van De Water, E. *Tongue Steerage in Frogs as a Repeatable Motor Neurological Program*. Poster Presentation, Experimental Biology, Washington, D.C. 1999.

ABSTRACT

A META-ANALYSIS OF TYPE I ERROR RATES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITH LOGISTIC REGRESSION AND MANTEL-HAENSZEL IN MONTE CARLO STUDIES

by

Eva C. Van De Water

Differential item functioning (DIF) occurs when individuals from different groups who have equal levels of a latent trait fail to earn commensurate scores on a testing instrument. Type I error occurs when DIF-detection methods result in unbiased items being excluded from the test while a Type II error occurs when biased items remain on the test after DIF-detection methods have been employed. Both errors create potential issues of injustice amongst examinees and can result in costly and protracted legal action. The purpose of this research was to evaluate two methods for detecting DIF: logistic regression (LR) and Mantel-Haenszel (MH).

To accomplish this, meta-analysis was employed to summarize Monte Carlo quantitative studies that used these methods in published and unpublished literature. The criteria employed for comparing these two methods were Type I error rates, the Type I error proportion, which was also the Type I error effect size measure, deviation scores, and power rates. Monte Carlo simulation studies meeting inclusion criteria, with typically 15 Type I error effect sizes per study, were compared to assess how the LR and MH statistical methods function to detect DIF.

Studied variables included DIF magnitude, nature of DIF (uniform or non-uniform), number of DIF items, and test length. I found that MH was better at Type I error control while LR was better at controlling Type II error. This study also provides a

valuable summary of existing DIF methods and a summary of the types of variables that have been manipulated in DIF simulation studies with LR and MH. Consequently, this meta-analysis can serve as a resource for practitioners to help them choose between LR and MH for DIF detection with regard to Type I and Type II error control, and can provide insight for parameter selection in the design of future Monte Carlo DIF studies.

META-ANALYSIS OF TYPE I ERROR RATES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITH LOGISTIC REGRESSION AND MANTEL-HAENSZEL IN MONTE CARLO STUDIES

by

Eva C. Van De Water

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Policy Studies
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2014

ACKNOWLEDGEMENTS

I owe thanks to many people for supporting me through the completion of my doctoral degree. First, I give thanks to my parents for instilling me with a love of learning from a young age. I thank my teachers all along the way for continuing to inspire me to learn. I give special thanks to my dissertation's chair, Dr. Curlette, for providing me with excellent advice and for being a wonderful mentor, to Dr. Oshima for her expertise on the subject of DIF and for answering my questions along the way, to Dr. Li for her for assistance and encouragement, to Dr. Alemdar for careful review and the perspective of a recent graduate, and to Dr. Sipe for the uplifting emails, and guidance through the meta-analysis process.

Without the support of my husband who not only believed in me but took care of our children while I pursued my dream of obtaining a doctoral degree, this path would not have been possible. I would also like to thank my employers and my colleagues for their support. Finally, I would like to thank my friends for words and gestures of support that inspired me to persist and provided me with much needed inspiration along the way.

I dedicate this dissertation to Nancy Van De Water, my mother-in-law, who was also a life-long learner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIXES

ABBREVIATIONS

1PL	One-parameter Logistic Model
2PL	Two-parameter Logistic Model
3PL	Three-parameter Logistic Model
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IRT	Item Response Theory
LR	Logistic Regression
MH	Mantel-Haenszel

CHAPTER 1: INTRODUCTION

Since the Civil Rights Era of the 1960s, testing fairness has been an important matter not only for testing and educational agencies, but the general populace as well (Hambleton, Swaminathan, & Rogers, 1991). Porter (2003, as cited in Kane, 2010) described *fairness* as a quality existing in the absence of bias, not only in democratic societies, but also in the field of measurement. Interpreted statistically, the term *bias* refers to the systematic under or over estimation of a parameter. For the statistically uninitiated, however, bias is synonymous with unfairness (de Ayala, 2009). Kane's (2010) view of fairness in testing is composed of two basic notions: the right of all people to be treated equally and the absence of bias.

Background of the Problem

Standardized testing is widely used in such arenas as college admissions, job placement, and job promotion. Considering the manner in which these test scores are used, it is essential that each item on a test functions the way it was designed to function (Kirshner & Guyatt, 1985). Imperfections plague even the most carefully developed testing instruments. A common flaw associated with standardized tests is differentially functioning items, which occurs when items on a test function differently for discrete groups having the same ability, such as males and females or majority versus minority groups. For example, if equal-ability members of these groups systematically interpret a question differently resulting in different answers, differential item functioning (DIF) is said to occur. Therefore, when it comes to standardized tests, developers and psychometricians wish to minimize DIF in their instruments (van de Vijver & Hambleton, 1996).

While it is desirable to identify items that contain DIF, it is highly undesirable to mark as compromised test items that function as desired; this latter condition is precisely what happens

when false positives occur, resulting in Type I error. One reason for inflated Type I error rates is the increase in statistical significance that accompanies large sample size (Jodoin & Gierl, 2001). As sample size increases, power tends to increase, thus leading to an increase in the number of DIF items being identified. Type I error occurs when DIF-detection methods result in unbiased items being excluded from the test while a Type II error occurs when biased items remain on the test after DIF-detection methods have been employed. Both errors create potential issues of injustice amongst examinees and can result in costly and protracted legal action.

Concern for equality in testing can be traced back to the 1960s (Camili & Shepard, 1994), when large mean differences in performance on test items were noticed between demographic groups. By 1972, Angoff reported that this concern for equality in testing began to shape into DIF. Today, a variety of statistical methods are used to detect DIF (Appendix A).

Methods to Detect DIF

Numerous parametric and nonparametric methods have been proposed for detecting DIF (Furlow, Ross, & Gagné, 2009; Rivas, Stark, & Chernyshenko, 2009; Woods, 2009), and many simulation studies have examined the performance of these methods to flag DIF items (Bolt & Cohen, 2001; Cohen & Kim, 1993; DeMars, 2009; Fidalgo, Hashimoto, Bartram, & Muniz, 2007; Fidalgo, Ferreres, & Muniz, 2004; Finch & French, 2008; French & Maller, 2007; Gómez-Benito, Hidalgo, & Padilla, 2009; Gonzalez-Roma, Hernandez, & Gómez-Benito, 2006; Güler & Penfield, 2009). Though a large body of research exists, the statistical technique of meta-analysis has not been used to summarize the Type I error of various statistical and item response theory (IRT) methods of DIF detection across simulation studies.

Even though the IRT method is not compared in this meta-analysis, it is significant because of the role it plays in simulations studies. In most studies, an IRT 1, 2, or 3 parameter

logistic (PL) model is used to generate items for the simulation study, therefore in this meta-analysis IRT model selection, e.g. 1PL, 2PL or 3PL, is relevant as a methodological study characteristic. The IRT model used by each included study is located in B. DIF magnitude and Nature of DIF methodological study characteristics are shown in Appendix C, while discrimination and difficulty parameters are shown in Appendix D. Number of replications per study are shown in Appendix E. Based on a review of available studies, logistic regression (LR) and Mantel-Haenszel (MH) are the methods most consistently reported in the literature for presentation of Type I error data and are often reported together. A comparison of LR, MH, and IRT methods is shown in Appendix F. In order to provide information to researchers concerning the efficacy of methods to measure DIF, studies that used two statistical methods and presented Type I error data in a way that allowed for calculations of proportions for each method were needed. Thus, LR and MH were the most logical choices for inclusion in the meta-analysis.

The empirical study of Bielinski and Davison (1998) and the simulation study of Monahan and Ankenmann (2005) confirmed that the effect of difference in ability variance between reference and focal groups is strong in DIF detection. The reference group is the larger group, often the non-minority group, for whom the item functions well. The focal group is usually the smaller group, frequently a minority group that experiences difficulty with a test item. This difficulty is not due to ability, but rather a result of the manner in which the item is written. However, Monahan and Ankenmann's study focused only on the MH chi-square test, whereas Bielinski and Davison's study focused only on the likelihood ratio test.

Use of Meta-Analysis and Systematic Review

Over four decades ago, Garvey and Griffith (1971) noted that scientists were being overloaded with information pertaining to their specialty. Methods needed to summarize the

existing body of literature 40 years ago are in even greater demand today. Systematic review and meta-analysis are two specific approaches to research synthesis. A query of the EBSCOhost search engine produced 203,439 citations for *meta-analysis* and 217,461 for *systematic review* between the years 1975 and 2013. Systematic review is a thorough search of existing literature, including published and grey literature, to collate pertinent data from articles on a specific topic. The resulting articles are then assimilated on comparable values in a consistent manner. Meta-analysis is a frequent though not mandatory quantitative component of a systematic review (Borenstein et al., 2009). However, it can function independently as a statistical technique, as in the case of this meta-analysis, since statistical comparison of substantive study characteristics is the focus of the study (Cooper, Hedges & Valentine, 2009). The goal of research synthesis is to "integrate empirical research for the purpose of creating generalizations" (Cooper, Hedges, & Valentine, 2009, p. 6). Though meta-analysis and systematic review have the shared goal of integrating empirical research for the purpose of generalization, meta-analysis is different because it summarizes data with statistical methods. Littell, Corcoran, and Pillai (2008) stated, "It analyzes trends and variations in research across studies, and [it] corrects for error and bias in a body of literature" (p. 2).

Problem Statement

Items that function differentially on a test for discrete groups having the same ability are called DIF items. A variety of statistical and IRT methods exist for the detection of DIF (Clauser & Mazor, 1998; French & Finch, 2013; Magis & Facon, 2012; Scott et al., 2010). Literature on the simulation and detection of DIF items is extensive and varied with regard to conditions manipulated and the statistical and IRT methods used to detect DIF in each study as shown in Appendix A. LR (a method to calculate odds ratios) and MH (a method to compare the

proportion of correct versus incorrect answers on a particular test item with membership of an examinee) are most consistently reported in the literature and are often reported together.

However, a search of the literature has revealed no statistics using meta-analysis that provide a quantitative summary of Type I error with LR and MH methods for detection of DIF items. A need exists for psychometricians and test developers to be able to compare DIF detection methods when deciding which DIF detection method they will use to analyze a particular testing instrument (Zappe, 2007). Therefore, this study seeks to summarize the efficacy of statistical methods for DIF detection in simulation studies using meta-analysis so researchers can have access to quantitative information about the manner in which LR and MH perform under different circumstances.

Purpose

This study reviewed articles using meta-analysis statistical techniques to provide a quantitative summary of Type I error with LR (Swaminathan & Rogers, 1990) and MH (Holland & Thayer, 1988; Mantel & Haenszel, 1959) methods for detection of DIF items. Thus, the aim of this study is to examine how the difference in Type I error for correct identification of differentially functioning items is affected by two commonly used DIF detection methods, LR and MH. The overarching goal is to summarize simulation studies to provide quantitatively based guidance for practitioners seeking a DIF detection method, e.g. LR or MH, tailored to their needs.

Research Questions

A meta-analysis of the methods used to detect differentially functioning test items using LR and MH is the focus of this study. The meta-analysis was conducted to answer the following research questions.

- Under various conditions in Monte Carlo computer simulations, how do the Type I error rates compare for LR and MH?
- How does each LR Type I error proportion & MH Type I error proportion compare to the accepted detection rate of 0.05? This 0.05 nominal Type I error detection rate indicates that incorrectly identifying a non-DIF item as DIF-containing for five percent of the non-DIF items on a particular simulated test is considered acceptable.
- How do the following substantive study characteristics affect Type I error effect size: impact, sample size, percentage of DIF and test length?
- How do Type II error rates compare for those studies displaying power data?

The fourth research question evolved in the following manner. After the ten included studies were screened for Type I error data, three of the included studies were found to display power data. Therefore, a comparison of power rates was included in research question four.

Theoretical Framework

This study is based on the concepts of DIF and meta-analysis. DIF analysis provides an indication of unexpected behavior of items on a test and is most often assessed using MH or LR. Meta-analysis allows researchers to statistically contrast and combine results from different studies to identify patterns among the results of these studies. Thus, this study uses meta-analysis to provide a quantitative summary of Type I errors in the LR and MH methods for detection of DIF items.

DIF Analysis

One axiom of modern test theory, according to Hambleton, Swaminathan, and Rogers (1991), is that observable traits, such as intelligence or verbal ability, can be predicted by an examinee's performance on a test. The estimation of observable traits, also referred to as latent

traits, is the goal of many modern tests. Tests are used for a variety of purposes from determining whether public schools students qualify for gifted education to testing the language skills of adults with an eye toward employment qualifications.

Because many tests are considered high-stakes tests, it is essential that each test is fair or unbiased. Measurement bias may be evaluated through either judgmental or statistical methods or a combination of the two methods (Zumbo, 1999). Use of judgmental methods to evaluate measurement bias uses a panel of experts who evaluate the test or test item from a human perspective. By contrast, statistical methods, such LR and MH, which are the focus of this study, provide a quantitative basis for evaluating bias. According to Zumbo (1999), "the technique called differential item functioning (DIF) analysis has become the new standard in test bias analysis" (p. 4).

The operational definition of DIF consists of three components that are pertinent to this study (Zumbo, 1999): 1) determining which subgroups will be analyzed, 2) deciding the amount of DIF magnitude that constitutes DIF, and 3) judging on what basis items will be reviewed (e.g., whether they favor the reference group or the focal, or both). First, a researcher who wishes to evaluate a test or test items for DIF must determine which subgroups will be analyzed. The most common subgroups are based on gender, race, culture, and language. Items containing universally inappropriate language or items that are biased to both subgroups are not considered DIF items since these items do not favor one group over another. Though analysis of DIF items for more than two subgroups is possible with LR, the 2x2 contingency table format of MH is not able to facilitate such an analysis (Clauser & Mazor, 1998). Therefore, only DIF analysis with two subgroups (e.g., male and female), reference and focal, with be investigated here.

Although DIF analysis falls on the statistical side of measurement bias, human input is still necessary. The second part of DIF's operational definition concerns DIF magnitude, a measure of how much DIF a particular item contains. Removal of all DIF-containing items from a test is generally impossible. Therefore, DIF magnitude, a numeric value associated with the amount of DIF present in a test item, provides quantitative information useful for comparison (Zumbo, 1999). Typically two steps are employed to reduce test bias. First, a statistical method provides a measure of DIF magnitude which is used to categorize the test items as DIF-containing or not DIF-containing. Next, a panel of experts evaluates the DIF items along a continuum to determine which items should be removed and which could potentially remain on the test (Camili & Shepard, 1994). In this study, simulation studies have been assessed with regard to their use of LR and MH statistical methods to evaluate DIF. The description of DIF given above is intended to provide context for the use of statistical methods in the detection of DIF as it related to decreasing the bias inherently present in most tests.

Meta-Analysis

Since the mid-1970s, meta-analytic methods have been widely used for research synthesis. Meta-analysis is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings" (Glass, 1976, p. 3). In the fields of education and medicine where multiple studies are conducted on focused research areas, meta-analysis is a valuable tool. When studies address the same topic, yet yield different results, meta-analysis provides a quantitative way to assess the differences among studies. Cooper et al. noted, "four key strengths of meta-analysis: parsimony, precision, objectivity, and "replicability" (2009, p. 511).

Though meta-analysis has many benefits, publishers and secondary researchers must maintain vigilance against bias. Bias that occurs due to the "selective publication of studies with a specific outcome, usually those which are statistically significant," (Ferguson & Brannick, 2011, p. 120) is called *publication bias*. This type of bias affects meta-analyses because avenues for acquiring published literature are often more convenient than are those available for the acquisition of unpublished literature. *Grey literature* (also referred to as gray or fugitive literature) is defined as "that which is produced on all levels of government, academics, business and industry in electronic and print formats not controlled by commercial publishers" (Auger, 1998, as cited in Cooper et al., 2009, p. 104). While this meta-analysis focuses mainly on the statistical comparison of LR and MH, the steps closely approximate those followed for a systematic review, with two exceptions: a comprehensive literature review and the use of multiple coders.

Definition of Terms

Background characteristics. Background characteristics are unchanging aspects of a study. Examples of background characteristics (Curlette & Canella, 1985), also called fixed study parameters, are author(s), publication date, and type of study (e.g., simulation study).

Differentially functioning items (DIF). DIF is a common flaw associated with standardized tests, occurring when items on a test function differently for discrete groups having the same ability, such as males and females or majority versus minority groups (Hambleton et al., 1991). If this item bias happens, one group will have an unfair advantage over the other. Psychometricians define DIF this way: "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton et al., 1991, p. 110).

DIF effect size. The DIF magnitude statistic provides a quantitative process for retention or removal of test items. While DIF studies frequently include an effect size, this DIF effect size (Wiberg, 2009) is used to measure DIF magnitude of specific tests and items; it is not an appropriate summary effect for meta-analysis and was not used in the current study.

DIF magnitude. DIF magnitude, which is expressed in many studies as DIF effect size, describes the amount of DIF present in a particular item.

DIF percentage. Contrasted with DIF magnitude, DIF percentage tells how many items on a test contain DIF, instead of the amount of DIF present in each item.

Effect size. The accepted benchmark used to compare outcome variables of studies on a common scale. Effect size has been embraced in recent years since it is robust with respect to sample size (Cooper & Hedges, 2009).

Type I error effect size. For the purposes of this study, Type I error effect size, calculated using Type I error, provides the common scale for comparison of study characteristics across studies. Type I error effect size was calculated as a proportion of incorrectly identified DIF items versus total number of non-DIF items on simulated tests taken by simulated examinees.

Focal. For the purposes of this study, the term *focal* was used to refer to the minority group (Holland, 1985).

Impact. The phenomenon occurring when one group earns higher scores on a test than another group as a result of true ability differences is called impact (Clauser & Mazor 1998).

Item characteristic curves (ICCs). The graphical performance of a particular test item can be shown with an ICC. The x-axis on the ICC shows the amount of the attribute being

measured (e.g., knowledge) while the y-axis shows the probability of answering the question correctly (DeVellis, 2011).

Item response theory (IRT). IRT is a model used to design, analyze, and/or score instruments that measure abilities, attitudes, or other variables. IRT is the preferred method for developing scales, especially when optimal decisions are demanded, as in standardized testing situations (Hambleton, Swaminathan & Rogers, 1991).

Logistic regression. Logistic regression uses probabilities to calculate odds ratios to determine if a test item is biased. Logistic regression can be divided into three categories based on the number of outcomes exhibited by the dependent variable. Binary logistic regression occurs, as in this study, when the observed variable has one of two possible outcomes (case or not a case). In logistic regression, the predictors or independent variables are used to predict the odds of being a case. The odds are calculated by dividing the probability that an outcome is a case by the probability that the outcome is not a case. If the two logistic regression curves overlap, a situation of no DIF is said to occur. If the curves are parallel yet do not overlap, uniform DIF is present, and if the curves cross, non-uniform DIF is present (Zumbo, 1999).

Mantel-Haenszel. The Mantel-Haenszel procedure uses a 2x2 contingency table to compare the proportion of correct versus incorrect answers on a particular test item with membership of an examinee in either the focal or the reference group. If DIF is not present, the proportion of correct to incorrect items for each group should be equal. MH is not a reliable method for identifying non-uniform DIF (Clauser & Mazor, 1998).

Meta-analysis. A meta-analysis uses a systematic review of research studies to contrast and combine results from different studies or, as in this case, it can be used as a statistical technique. Using statistical techniques, meta-analysis identifies patterns among the selected

studies, finds possible sources of disagreement among those results, and highlights other relationships that may be of interest to researchers (Cooper, 2004).

Methodological study characteristics. Methodological study characteristics are tied to the steps carried out during primary research (Curlette & Canella, 1985). Examples of methodological study characteristics include: generating model (e.g., 2PL or 3PL), item parameters (e.g., *a* or *b*) and number of replications.

Reference. For the purposes of this study, the term *reference* is used to refer to the majority group (Holland, 1985).

Substantive study characteristics. Substantive study characteristics have the potential to affect the outcome variable (Curlette & Cannella, 1985). Examples of substantive study characteristics include: impact, sample size, percentage of DIF, and test length.

Systematic review. A systematic review begins with a search of the literature using specific rules and is followed by inclusion or exclusion of studies according to clear criteria. Meta-analysis is often included as a quantitative component of systematic reviews (Littell, Corcoran & Pillai, 2008).

Type I error. Type I error, which is a false positive, occurs when DIF-detection methods result in unbiased items being excluded from a test (Jodoin & Gierl, 2001).

Type II error. When biased items remain on the test after DIF-detection methods have been employed, Type II error or a false negative has occurred, creating potential issues of injustice amongst examinees that can tend toward litigation (Jodoin & Gierl, 2001).

Research Goals

This dissertation seeks to provide those who develop and use standardized tests with a quantitative summary of LR and MH methods for detecting DIF. A literature search was

conducted to find studies that simulated DIF. Only simulation studies that met the following inclusion criteria were used: (a) employed LR and MH to detect differentially functioning items, (b) constituted a simulation study, (c) contained Type I error data either in summary form or by condition, and (d) reported between 1975 and 2013.

Studies were excluded if they (a) only used real data; (b) did not contain Type I error data, including studies containing data (e.g., means and standard deviation) that possibly has been converted to effect size data; (c) examined either LR or MH, but not both; (d) presented Type I error results, but not the raw Type I error data needed to calculate the Type I error effect size or; (e) were not available in English. The ten simulation studies included in the meta-analysis are listed in Appendix G.

Assumptions

For this study, I present the following assumptions:

- Simulation studies are accurate representations of the real-word situations they attempt to simulate.
- Simulation studies are carried out according to the methods described therein.
- Type I error is calculated properly in each of the articles reviewed.
- Meta-analysis is the proper tool for quantitatively summarizing outcomes of simulation studies focusing on LR and MH methods for DIF detection.

Limitations

The present study has the following limitations:

- Only 10 studies met inclusion criteria.
- Each of the included studies was published.
- The study depends on existing research for accurate data.

- The study uses simulation studies consisting of multiple conditions, which can present challenges from the analysis perspective.
- Data was extracted independently.

Summary

Fairness as a quality is important in the field of measurement. Each item on a test must function in the way it was designed. However, even the most carefully developed testing instruments can be plagued with imperfections. A common flaw associated with standardized tests is DIF, which occurs when items on a test function differently for discrete groups having the same ability. Finding statistical methods that can detect DIF while minimizing Type I and Type II errors is important for psychometricians. Therefore, this study was designed to use the statistical methods of meta-analysis to examine how the difference in Type I error for correct identification of differentially functioning items is affected by two commonly used DIF detection methods, LR and MH.

CHAPTER 2: REVIEW OF THE LITERATURE

This literature review provides a background of research in the field of DIF and explains in abbreviated form each of the statistical and IRT methods pertinent to the meta-analysis. At present, there are multiple ways to assess DIF, and no summative research such as meta-analysis has been conducted to assess the success of different methods and statistical measures used to identify DIF (Guilera, Gómez-Benito, & Hidalgo, 2010).

History of Meta-Analysis

Research synthesis may be applied to any discipline containing documents whose contents can be summarized for subsequent use. For example, motivated to address discrepancies in the treatment of scurvy and typhoid, respectively, James Lind, in the 1700s, and Karl Pearson, in 1904, set about the task of analyzing primary documents in an effort to summarize existing knowledge about each disease. In 1907, Joseph Goldberger conducted a statistical synthesis of typhoid that implemented four steps integral to meta-analysis: review of the literature, use of specific criteria to select studies, abstraction of the data, and statistical analysis of the abstracted data (Chalmers, Hedges, & Cooper, 2002). At a 1976 presidential address highlighting the need for a "better synthesis of research results," Glass introduced the term meta-analysis (Chalmers et al., 2002). By 2004, meta-analysis was being used in finance, marketing, sociology, wildlife management, and economics, in addition to education and medicine (Hunter & Schmidt, 2004).

Historically, meta-analysis preceded the implementation of specific steps for reducing bias (Chalmers et al., 2002). Regardless of the discipline, the need to control for bias is present in primary and secondary research. The six type of bias most commonly found in meta-analysis are biases of: publication, databases, citations, multiple publications, inclusion criteria, and provision of data (Egger, Davey Smith, Schneider & Minder, 1997).

Price's (1965) view that research syntheses serve to "replace those papers that have been lost from sight behind the research front" (p. 513) seemed to err on the side of inclusion. In the same vein, Glass (1978) embraced the unstandardized nature of education research instead of forcing it into a preexisting framework; he supported the inclusion of a variety of studies regardless of the rigor of their research design. In his opinion, the importance of study design is diminished when the studies' findings have a small covariance when compared with similar studies. Instead of weeding out studies with imperfect research design from the outset, Glass preferred to use crosstabs or other quantitative analyses to reveal differences in research methodology. Indeed, his view may protect against bias related to inclusion criteria since statistical methods are used to make decisions regarding exclusion and inclusion of studies in lieu of researchers' opinions concerning quality of methodology and perceived differences in studies.

Meta-Analysis as Research Methodology

The introduction of null hypothesis significance testing by R.A. Fisher in 1932 marked the beginning of a long line of statistical methods that have been used in an attempt to summarize the literature quantitatively (Chalmers et al., 2002; Hunter & Schmidt, 2004; Sipe & Curlette, 1997). Light and Pillemer (1984) viewed positive tests for statistical significance simply as a first step in demonstrating the effectiveness of research methods, and by the early 2000's the popularity of significance testing for meta-analytic comparison has decreased due to its susceptibility to sample size (Cooper, 2004; Cooper et al., 2009; Hunter & Schmidt, 2004). Prior to the advent of meta-analysis, studies were compared using contingency tables and the presence or absence of statistical significance (Glass, 1978). Since statistical significance testing provides information solely on the probability that obtained results are due to chance, an actual

mathematical representation of the difference in treatment effects for groups of interest adds relevant information. *Effect size* provides that information. Rosenthal (1991) describes effect size as the size of the relationship between any two variables. According to Rosenthal's definition this meta-analysis uses four different effect sizes to compare MH and LR statistical methods for the detection of DIF: Type I error rates, deviation of Type I error rates from the nominal .05 level, Type I error effect size calculated as *d'*, and power rates. In 1999, the Wilkinson Task Force on Statistical Inference recommended that effect size always be reported and emphasized the need for meta-analysis in future research and the importance of effect size to meta-analysis. Huberty (1972) reported, "Depending on how one defines effect size, it may be claimed that its history started around 1940, or about 100 years prior to that" (p. 227). In 2001, Elmore (as cited in Huberty, 2002) counted 61 effect size choices.

Glass, McGaw, and Smith (1981) stated that, "Meta-analysis is an approach to quantitative synthesis of research studies which uses many techniques of measurement and statistical analysis to integrate numerous and diverse findings of research studies" (p. 8). According to Glass (1976), data analysis consists of three levels: primary analysis examines original data; secondary analysis answers new questions with old data, for example, improving statistical techniques; and meta-analysis is the analysis of analyses. Glass (1976) coined the term *meta-analysis*, but Light and Smith (1971) implemented a process they called *the cluster effect* a few years earlier. In reference to summarizing education studies, Light and Smith stated, "Progress will only become [possible] when we are able to pool, in a systematic manner, the original data from the studies" (p. 443). Wolf (1986) reported that strengths of meta-analysis included the following:

- Studies are summarized effectively.

- Studies are analyzed with statistical methods, which often results in stronger conclusions than literary reviews.
- A variety of studies are included, even ones that have weak research designs.
- Gaps in the literature are highlighted to provide new directions for further research.
- Mediating or interactional relationships or trends that cannot be hypothesized or tested in individual studies are discovered.
- Outliers are identified that may lead to increased understanding and new hypotheses.

Including a variety of studies enhances the meta-analysis. Even if a number of studies considered poor in technique are included, they may well add to the richness of the final data, particularly if the studies are not weak in the same areas (Glass, 1976). Therefore, Glass incorporated not only studies with strengths in implementation, but also those with clever research design. Though critics, (Eyseneck, 1978; Light & Pillemer, 1948), of Glass' inclusion of studies lacking the proper proportion of similar characteristics suggested he was comparing *apples and oranges*, he replied by reminding them that apples and oranges are both fruit and also inquired as to the purpose of comparison when studies are already quite similar. Glass tended to err on the side of including studies that may seem different to enrich the overall summary.

Though it does have many advantages, meta-analysis has disadvantages as well. Critics contend that vastly different studies should not be compared (Borenstein, Hedges, Higgins, & Rothstein, 2009; Light & Pillemer, 1984), but Glass (1978) maintained there would be no basis for meta-analysis if one only compared studies that were the same. The leniency meta-analysis shows for including different studies extends to poor studies as well. The inclusion of poor studies alongside good ones is endorsed by Glass (1976) as well as Hunter and Schmidt (2004). One of the strengths of meta-analysis is the ease with which studies that use varying substantial

and methodological characteristics can be summarized. Studies may differ in a variety of ways (e.g., sample size, publication type, and level of rigor with respect to methods). Quantitative assessment of studies with differing characteristics allows the effect sizes of the studies to be calculated and therefore known. If effect sizes between studies vary greatly, researchers may continue the analysis in an effort to detect the presence of moderator variables, which could be used to divide the studies into subsets (Hunter & Schmidt, 2004). Light and Pillemer (1984), recommend examination of effect size measures to determine whether the differences between studies is attributable to sampling error, meaning chance, or if it could be the result of true differences in treatment effects. Heterogeneity is the term used to describe this comparison of studies. Though heterogeneity is often discussed quantitatively, it is possible for studies with similar effect sizes to manifest qualitative differences that belie meaningful comparison. Therefore, comprehensive comparison of studies utilizing both quantitative and qualitative methods is indicated (Light & Pillemer, 1984).

Use of Meta-Analysis to Summarize DIF Detection Methods

The issue of DIF has been around since the mid-1970s, and since that time numerous methods have emerged to detect DIF (Hambleton et al., 1991). Some methods are based on classical test theory while others fall under the IRT umbrella. A wide array of DIF detection methods are available to researchers as shown in Appendix A. DIF detection methods can be classified as parametric or nonparametric, for dichotomously or polytomously scored items, for two groups or three or more groups, and as inclusive or exclusive of non-uniform DIF. MH is an example of a contingency table method, while LR belongs to a family of nested model methods, and IRT methods use likelihood ratios.

Each level of classification introduces new possibilities for the manner in which the variety of DIF detection methods may function. Meta-analysis provides a quantitative filter with which to sort the DIF detection methods as well as a means to detail the circumstances for the use of each one. Though meta-analysis researchers can chose from a variety of quantitative snapshots to summarize studies, some statistics are more appealing than others. The appeal of a particular statistic is dependent on several factors. The widespread use of significant *p*-values as a precursor to publication makes them easy to find in most publications, although the sensitivity of these values to sample size diminishes their value as a measure of comparison (Borenstein et al., 2009; Coulter-Kern et al., 2009). Power is equally susceptible to sample size; increasing its strength with increasing sample size (Cooper, Hedges, & Valentine, 2009).

In meta-analysis, effect size is the gold standard for comparing various studies using the same benchmark (Glass, 1978). In DIF studies, however, the term *effect size* takes on a different meaning. Here, DIF effect size is synonymous with DIF magnitude and is assigned to each DIF-containing item on a particular test. In this context, DIF effect size or magnitude gives test makers information concerning which items, of the ones identified as displaying DIF, are the most problematic. This DIF magnitude statistic provides a quantitative process for retention or removal of test items. The terminology surrounding summary effects of DIF can be confusing. Many DIF studies do include an effect size, but this DIF effect size (Wiberg, 2009) is used to measure DIF magnitude of specific tests and items; it is not an appropriate summary effect for meta-analysis and was not used in the current study. Instead this study utilizes Type I error effect size, which is a proportion of incorrectly identified DIF items divided by the total number of items on the test, Type I error rates, deviation scores of Type I error rates from the .05 nominal

level, and power rates to compare the efficacy of LR and MH statistical methods in evaluating DIF.

Type I error is a recurring statistic that is an appropriate indicator of the success of the method for detecting DIF. It is particularly appropriate as a summary effect, since a major issue with both LR and the MH procedure is inflated Type I error (Penfield, 2009). Certain circumstances increase the possibility of Type I error inflation, including the existence of both equal and unequal ability distributions (Narayanan & Swaminathan, 1996).

Differences in the formulas used and the steps employed for DIF detection demonstrate strengths and weaknesses of the various methods. Some, like MH, are more efficient and cost effective (Clauser & Mazor, 1998; Penfield, 2001; Wang & Su, 2004), while others, like the IRT methods, are more comprehensive, and come with an increased cost of time and money (de Ayala et al., 2009; Hambleton et al., 1991; Raju, van der Linden, & Fleer, 1995). Other considerations psychometricians may wish to consider including the types of DIF they want to identify and what steps can be taken once DIF is identified. Test bias can take on a litigious face especially if minorities are the disadvantaged group, or focal group, in the presence (Clauser & Mazor, 1998) or absence (Linn & Darsgow, 1987) of DIF. For this reason, the accurate detection of DIF and appropriate adaptations to the test if DIF is detected are paramount to test makers.

Fairness in Testing

Bias on standardized tests has been addressed in the literature (Bradbury, 2011; Cole, 1973, 1981; Nankervis, 2011). Cole (1973) presented selection bias and selection fairness as different sides of the same coin. Of the six models of fairness presented by Cole (1973), Darlington's (1971) model allows for the insertion of various cultural groups (Cole, 1973). Darlington's definition of test fairness (as cited in Newman, Hanges, & Outtz, 2007) emphasized

that race should not affect the chance that equal-ability examinees have of being selected for inclusion in a particular group based on test scores. Validity and fairness are connected because an unfair instrument that "systematically misrepresents the standing of some individuals or some groups of individuals on the construct being measured or that tends to make inappropriate decisions for individuals or groups is, to that extent not valid for interpretation or use" (Kane, 2010, p. 181).

In educational assessment, one concern is that specific groups of examinees defined by gender, ethnic, or other types of group membership earn lower scores than other groups (Greatorex & Bell, 2004). Though groups of people can differ in many ways, DIF analysis can only be applied to groups with manifest differences like race or gender. Splitting one of these groups, such as gender, into male and female segments, creates the focal and reference groups whose test responses are compared when evaluating an instrument for DIF. Here the focal group, often the smaller minority group, is the group being examined for DIF, while the reference group, often the larger majority group, is the comparison group (de Ayala et al., 2002).

The phenomenon occurring when one group earns higher scores on a test than another group as a result of true ability differences is called *impact* (Clauser & Mazor, 1998). Alternatively, a test may favor one group over the other due to bias. On a biased test, two equal-ability groups, such as males and females, do not have an equal opportunity to earn a score on the test commensurate with their ability. If this happens, one group will have an unfair advantage over the other. In the testing industry, this is known as DIF. Psychometricians define DIF this way: "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton et al., 1991, p. 110). This definition assumes the current practice of checking for the presence of DIF at the item level.

Though an item must be flagged for DIF in order to be considered biased, not every DIF item is actually biased. Therefore, evaluation of test items for DIF is a two-step process beginning with empirical analysis and progressing to qualitative inspection by a panel of experts (de Ayala et al., 2002). Expert advice is crucial because creating standardized tests is already expensive and test developers are therefore reticent to incur the added expense of removing items unless it is absolutely necessary.

Introduction of the Empirical Methods for DIF-Detection

A wide array of DIF detection methods is available, and it is important to demonstrate a test is free of bias. This meta-analysis focuses on the primary empirical step of DIF detection, specifically identifying whether LR or MH is the appropriate method to locate DIF in a variety of situations. DIF detection methods bring their own advantages and disadvantages to the analysis. Each of the statistical methods for DIF detection compares the performance of two groups on a studied item.

Before comparisons can be made, the two groups must be matched on a measure of ability (Clauser & Mazor, 1998). IRT methods use between-group differences on item parameters to model DIF data but require large sample sizes (Hambleton et al., 1991). In fact, two of the IRT parameters: *a* (discrimination) and *b* (difficulty) are used in many of the simulation studies in this meta-analysis to generate DIF items (Jodoin & Gierl, 2001).

Since DIF is represented visually using *item characteristic curves* (ICCs), the introduction of a second operational definition may be helpful: "An item shows DIF if the [ICCs] across different subgroups are not identical. Conversely, an item does not show DIF if the [ICCs] across different subgroups are identical" (Hambleton et al., 1991, p. 110). IRT item characteristic curves can be used to visually depict the latent trait, e.g. ability, of an examinee plotted against

the probability of the examinee answering the item correctly. The curves typically take on an 'S' shape with an asymptote at either end of the latent trait continuum depicted on the x-axis (Osterlind & Everson, 2009). This latent trait or theta is then plotted on the x-axis while the probability of a correct response (P) is plotted on the y-axis. ICCs have three possible components or parameters: discrimination (*a* parameter), difficulty (*b* parameter), and pseudo-guessing (*c* parameter). The discrimination (*a)* parameter, determines the slope of the ICC; a curve that is closer to vertical does not discriminate well between examinees of varying ability while a curve with a more horizontal shape does discriminate well among examinees with different theta, or ability, values. The *b* parameter, which indicates item difficulty, is present in all IRT models and ICCs (Harris, 1989). As the value of the *b* parameter changes, the position of the ICC moves along the x-axis. A curve situated farther to the left represents an easier question, and a curve situated farther the right indicates a more difficult question. Therefore, higher theta values are associated with higher ability levels and lower theta values with lower ability. The pseudo-guessing (*c*) parameter shows the likelihood of an examinee answering a question correctly by simply guessing. Often the *c* parameter is set to 0.20 to indicate the probability of an examinee answering the question correctly by guessing on a multiple choice test with five answer choices (Hambleton et al., 1991).

One-Parameter Model. Item characteristic curves which display only *b* parameter changes, as shown in Figure 1, are referred to as one-parameter models (Hambleton et al., 1991). The one-parameter logistic model (1PL) assumes that the only item characteristic affecting examinee performance is item difficulty.

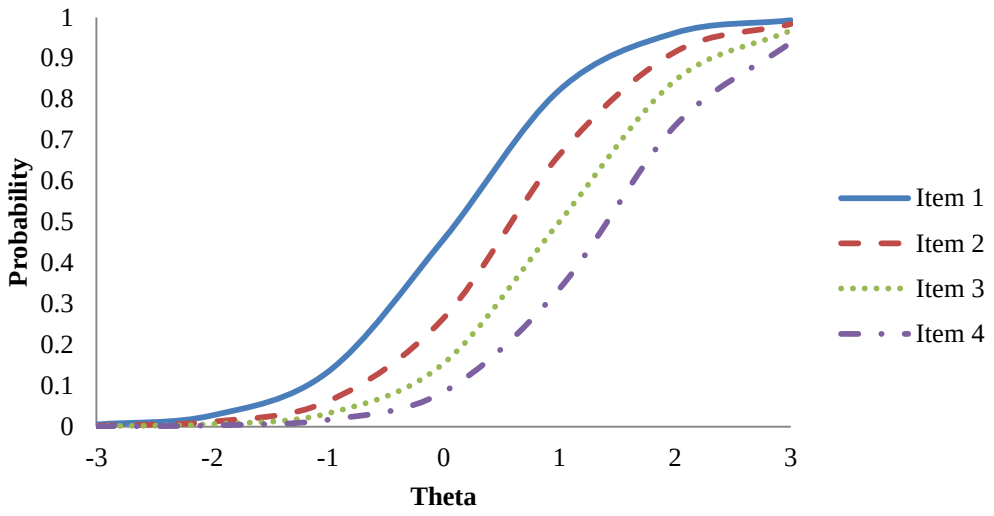


Figure 1. One-parameter model item characteristic curve showing four items with varying difficulty (*b*) parameter values.

Because the 1PL model does not include the *c* parameter (pseudo-guessing), the lower asymptote is 0, meaning that examinees of very low ability will have zero probability to answer the item correctly (Hambleton et al., 1991). Since the discrimination parameter (*a*) is held constant, the slope of the ICC is held constant. Here, item 1 is the easiest item and item 4 is the most difficult item. When used with dichotomous data, the 1PL model, sometimes referred to as the Rasch Model, has three distinct advantages over the two-parameter logistic model (2PL) and three-parameter logistic model (3PL): total test score can be used to estimate theta level (ability), the number of examinees answering a question correctly can be used to estimate the *b* parameter (difficulty), and examinees having the same raw score will have the same theta level (Harris,1989; Osterlind & Everson, 2009). Changes in the 1PL occur when a curve with a constant slope shifts to different points on the x-axis, demonstrating the varying difficulty of items.

Two-Parameter Model. A model which includes the *a* and *b* parameters representing discrimination and difficulty, respectively, is called the two-parameter model (2PL).

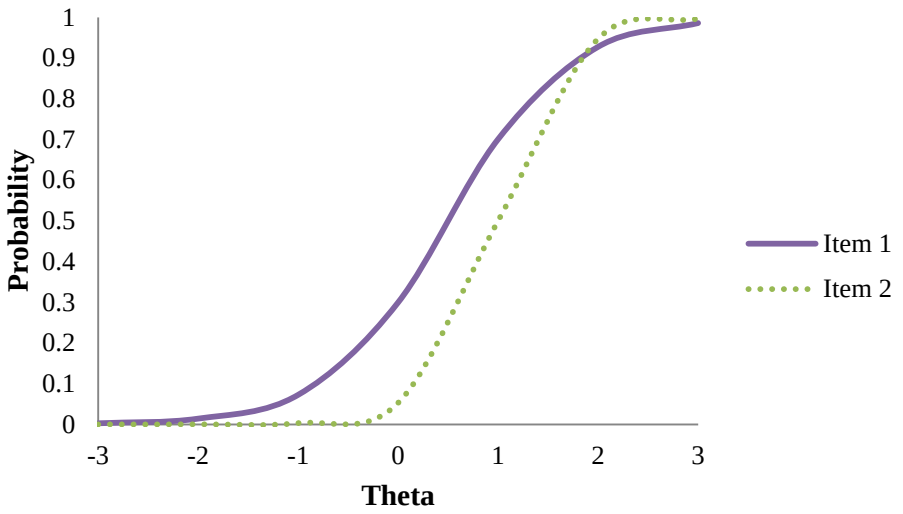


Figure 2. Two-parameter model item characteristic curve: difficulty (*b*) and discrimination (*a*) parameters vary while pseudo-guessing parameter (*c*) is held constant.

Here, *discrimination* indicates the ability of the item to distinguish between examinees of varying ability levels. The closer the slope over a range is to vertical, the better ability the item has to discriminate. A curve with a gentler slope will be less useful when discriminating between examinees of varying ability levels. Here item 2 which has a steeper slope is more discriminating. Because the slopes of the curves are unequal, the curves cross. In Figure 2, skill level is labeled *theta* and falls between -3 and 3 on the *x*-axis. A theta value of -3 indicates an examinee at the lowest skill level, while a theta level of 3 indicates an examine at the highest skill level. Therefore, the 2PL model combines the slope (*a* parameter) with the position of the curve on the x-axis (*b* parameter) allowing the ICC to display not only item discrimination between examinees, but also item difficulty.

Three-Parameter Model. When ICCs depict a three-parameter logistic model (3PL), values for each of three parameters, *a*, *b* and *c*, are expected to influence the examinees'

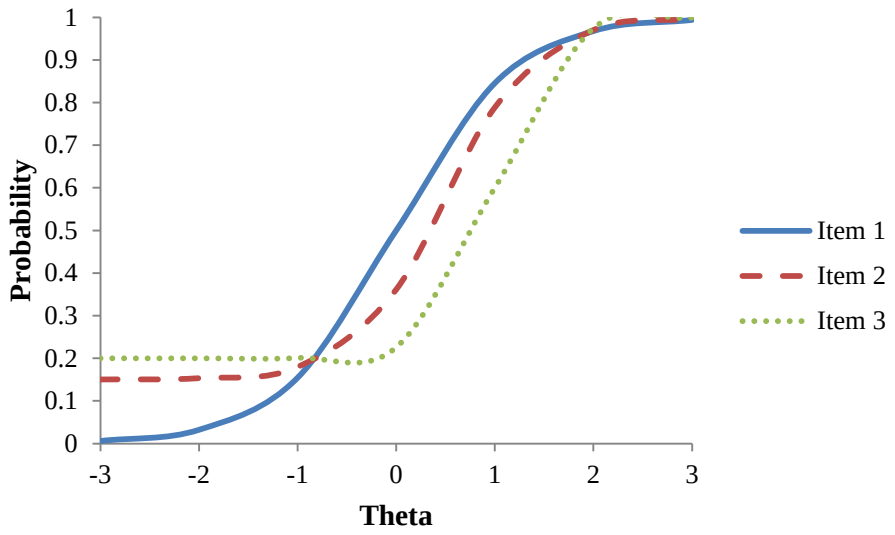


Figure 3. Three-parameter model item characteristic curves intersect with the y-axis at different values for each of the three items representing varying values of the pseudo-guessing (*c*) parameter.

performance on the item. The *c* parameter, or pseudo-guessing parameter, is the lower asymptote and indicates the likelihood that an examinee will answer the item correctly if he or she simply hazards a guess (Hambleton et al., 1991). On a multiple-choice test with five answer choices, the probability of an examinee achieving a correct answer by guessing is 20%, shown by item 3 in Figure 3. For this reason, the *c* parameter is often set to 0.20 in simulation studies; studies examinees of the lowest ability (theta = -3) would not be expected to answer correctly. The included in this meta-analysis follow this convention. Item 1 in Figure 3 shows an item which methods (Swaminathan & Rogers, 1990). Invariance means that the items and tests function independently of the examinees.

IRT methods work by empirically examining differences in how test items function for reference and focal groups. For example, test items using technical hunting terms would likely be more difficult for women to answer than for men, even if the two groups have equal ability levels. According to Hambleton et al. (1991), one of the positive attributes of IRT methods is

that they provide "a unified framework for conceptualizing and investigating bias at the item level" (p. 8). Another advantage of IRT models is that they "do take into account the continuous nature of ability when comparing the performance of groups of examinees" (Swaminathan & Rogers, 1990, p. 362). One drawback of IRT models is the large sample size required for analysis. Typically minimum sample size is 500 per group for the Rasch, or 1PL, model. To implement the 2PL or 3PL sample sizes of 800 to 1,000 per group will be necessary according to Hambleton et al. (1991). Other drawbacks of IRT models include their sensitivity to model fit and the expense associated with implementation of the models.

IRT models can be used for dichotomous items (e.g., with two answer choices such as *true* and *false*) or polytomous items (e.g., with more than two answer choices). The drawbacks of IRT include the unidimensionality assumption, which presumes that "one dominant ability" (p. 10) is sufficient to explain examinee performance, and the need for a large sample size (Hambleton et al., 1991). The matching variable for IRT models is a measure of latent ability instead of a test score, which is used by the MH and LR procedures (Clauser & Mazor, 1998).

When testing for DIF, the null hypothesis for the 1P is that the difficulty item parameter is the same for the reference and focal groups. For the 2PL and 3PL models, the null hypothesis states that the ICCs for the reference and focal groups are the same (Clauser & Mazor, 1998). The fundamental building block of IRT is the ICC, which links the latent ability to the probability that a randomly drawn examinee of a given ability will answer the item correctly (Zajonc, 2009).

Uniform DIF and non-uniform DIF. DIF can be uniform, "meaning that one group of examinees is consistently unduly disadvantaged by the item under investigation, or non-uniform or crossing, meaning that the relationship reverses at some point along the scale" (Robitzsch &

Rupp, 2009, p. 23). An item characteristic curve depicting uniform DIF is shown in Figure 4. In other words, uniform DIF occurs when only the difficulty parameter differs across groups, and non-uniform DIF occurs when an interaction between ability level and group membership causes the item discrimination parameter to differ across groups at every ability level (Chan, 2000). Graphs of uniform DIF, such as Figure 4, show parallel curves; graphs depicting

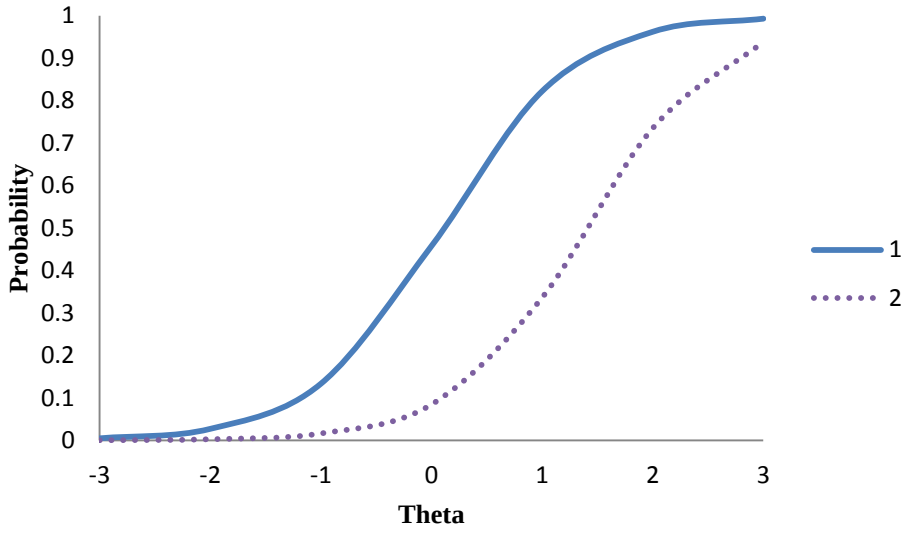


Figure 4. Item characteristic curves displaying varying difficulty (*b)* parameters, but constant discrimination (*a*) parameters illustrating uniform DIF.

non-uniform DIF, such as Figure 5, show the intersection of the two curves at the point where the advantage of higher scores switches from one group to the other.

When non-uniform DIF occurs, the reference group answers questions correctly at one range of ability but answers items incorrectly at another ability level. Simultaneously, the focal group answers questions incorrectly at one range of ability yet answers questions correctly at another ability level (Hambleton et. al., 1991). Such a phenomenon is exemplified by Figure 5.

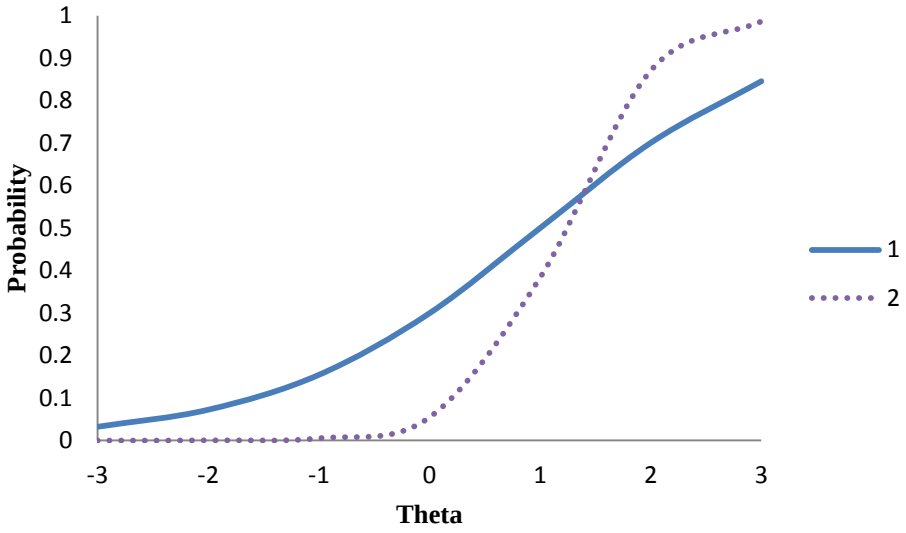


Figure 5. Item characteristic curves with equal difficulty (*b*) parameters but varying discrimination (*a)* parameters illustrating non-uniform DIF.

In figure 5 examinees with theta values less than 1 are more likely to answer item 1correctly, but at theta values above 1 examinees are more likely to answer item 2 correctly. Kanjee (2007) differentiated between uniform and non-uniform DIF by stating, "Uniform DIF occurs when there is no interaction between ability level and group membership, while for non-uniform DIF there is an interaction between ability levels and group membership" (p. 52). Non-uniform DIF can be further subdivided into symmetrical and asymmetrical DIF categories. Symmetrical non-uniform DIF occurs when only the discrimination parameter is modified, while asymmetrical non-uniform DIF occurs when there are differences in the difficulty and discrimination parameters (Hidalgo & Lopez-Pina, 2004).

Using IRT methods to detect non-uniform DIF. Hambleton and Rogers (1989) found that the IRT-based area method performed well with respect to detection of non-uniform DIF. The variation with the IRT method occurred because results were dependent on the interval selected (Hambleton & Rogers, 1989). They pointed out three drawbacks of the IRT methods, particularly the 3PL: high cost, large sample size requirements, and poor parameter estimates.

Statistical Methods for the Detection of DIF

Many statistical and IRT methods exist for the identification and evaluation of DIF (Kim, Cohen, Alagnoz, & Kim, 2007). In the following section, the LR, MH, and IRT methods with formulas are reviewed. A list of methods for detection of DIF can be found in Appendix A.

Logistic Regression

Logistic regression (LR) is popular with many statisticians for its ease of use with common statistical software (Zumbo, 1999), its ability to identify uniform and non-uniform DIF simultaneously, and its ability to handle dichotomous and polytomous items (Gómez-Benito et al., 2009). Zumbo reported that, "Logistic regression is based on statistical modeling of the probability of responding correctly to an item by group membership and a criterion or conditioning variable" (p. 29). LR was originally proposed by Swaminathan and Rogers (1990) to detect uniform and non-uniform DIF in dichotomous items. Swaminathan and Rogers also noted that LR can be easily expanded to accommodate two or more ability estimates.

LR is a bridge between contingency table methods and treats total score as a continuous variable. Depending on the model chosen, researchers can test for uniform DIF only, for uniform DIF or non-uniform DIF, or compare the models' fit to test for uniform DIF and non-uniform DIF simultaneously. LR is superior to MH when testing for non-uniform DIF (Swaminathan & Rogers, 1990).

LR is popular with many statisticians for its ease of use with common statistical software (Zumbo, 1999), its ability to identify uniform and non-uniform DIF simultaneously, and its ability to handle dichotomous and polytomous items (Gómez-Benito et al., 2009). Zumbo reported that, "Logistic regression is based on statistical modeling of the probability of responding correctly to an item by group membership and a criterion or conditioning variable"

(p. 29). LR was originally proposed by Swaminathan and Rogers (1990) to detect uniform and non-uniform DIF in dichotomous items. Swaminathan and Rogers also noted that LR can be easily expanded to accommodate two or more ability estimates. The nested nature of LR allows it to handle multidimensional data readily (Mazor et al., 1995, as cited in Hidalgo & Lopez-Pina, 2004). The general equation for LR can be written:

$$P\left(y = \frac{1}{X}\right) = \frac{e^z}{1+ e^z} \quad (1)$$

> "Where *y* is the answer to the item, P(y = 1/x) is the conditional probability of obtaining a correct answer given *X*, and *z* represents the linear combination of the predictor variables" (Gómez-Benito et al., 2009, p. 18).

In DIF analysis, the LR equation can be written:

$$Y' = a + b_1X_1 + b_2X_2 + b_3 X_1* X_2 \quad . \quad (2)$$

Here $b_1 X_1$ is the ability level of the subject measured by total test score, b_2X_2 is the group variable (reference or focal) and $b_3 X_1* X_2$ is the product of the ability and group variables (interaction variable). The intercept parameter is *a*, b_1 is the parameter corresponding to ability difference in performance on the item, b_2 is the parameter corresponding to group difference in performance on the item, and b_3 is parameter

$$\ln\left[\frac{p_i}{(1-p_i)}\right] = a + b_1X_1 + b_2X_2 + b_3 X_1* X_2 \quad (3)$$

corresponding to the interaction between group and ability (Zumbo, 1999). Table 1 contains a summary of the nested model formula variable meanings for LR and DIF criteria according to Gómez-Benito et al. (2009).

Table 1

DIF Criteria

No DIF	Uniform DIF	Non-uniform DIF
		$b_3 \neq 0$
$b_2 = b_3 = 0$	$b_2 \neq 0$ & $b_3 = 0$	($b_2 \neq 0$ or $b_2 = 0$)
No group difference & no interaction between group & ability	Difference between groups but no interaction difference	There is an interaction between group & ability

Note. Adapted from "Efficacy of Effect Size Measures in Logistic Regression: As application for detecting DIF," by J. Gómez-Benito, M. D. Hidalgo, and J. L. Padilla, 2009, *Methodology*, *5* (1), p. 19.

Thomas and Zumbo (1998) pointed out that Swaminathan and Rogers' (1990) equation for the probability of a correct response for DIF detection (Equation 1) is "nonlinear with respect to the odds or probability" (p. 24). Zumbo therefore used Equation 2 where Y' is a natural log of the odds ratio and where p is the proportion of individuals that endorse the item in the direction of the latent variable. One can then test the 2-degrees of freedom chi-square test for both uniform and non-uniform DIF (Zumbo, 1999, pp. 23-24). According to Zumbo (1999), advantages of using LR over other DIF methods such as MH are (a) one need not categorize a continuous criterion variable; (b) one can model uniform and/or non-uniform DIF (Swaminathan, 1994, as

Table 2

Suggested Regression Procedures for the Identification of DIF

Wald Statistic	Comparison of nested models	Overall consensus
Indicates significance of regression coefficients	Relative fit indicates whether DIF exists and the type	Model comparison superior

Note. Adapted from "Efficacy of Effect Size Measures in Logistic Regression: As application for detecting DIF," by J. Gómez-Benito, M. D. Hidalgo, and J. L. Padilla, 2009, *Methodology*, *5*(1), p. 19.

cited in Zumbo, 1999); and (c) one can generalize the binary LR model for use with ordinal item scores. (Zumbo, 1999, p. 23).

Comparison of nested models in LR. According to Gómez-Benito et al. (2009), "DIF can be detected by either using the Wald Statistic, which indicates the significance of regression coefficients or by the comparison of nested models" (p. 19). They recommend the use of nested models based on the results of their simulation study, as shown in Table 2.

The nested study approach, compared in Appendix H, requires a three-step process for evaluating the model using equation 2. Using Zumbo's (1999) equation, the first phase introduces the total test score (X_1) into the base equation. The second phase introduces the grouping variable (X_2) into the equation to test for uniform DIF. The final phase incorporates the interaction variable ($X_1 * X_2$) into the equation. To complete the test for uniform and non-uniform DIF the likelihood functions of the three models are compared. If the LR curves are the same for the two groups, no DIF is present (Swaminathan & Rogers, 1990). Zumbo tested for DIF comparing models using a likelihood function to calculate R^2, while Swaminathan and Rogers (1990) conducted the same comparison using slopes and intercepts from the nested regression equation to test for uniform and non-uniform DIF (Table 3). Following Zumbo's method, to test for uniform DIF the Model 1 R^2 is subtracted from the Model 2 R^2. If the result is zero, there is no DIF; if the result is not zero, there is uniform DIF.

Swaminathan and Rogers (1990) performed a similar calculation using the slopes and intercepts. If the difference in the slopes is zero, but the difference in the intercepts is not zero, uniform DIF occurs because the curves are parallel but not overlapping (which would indicate no DIF). To test for non-uniform DIF, Zumbo (1999) subtracted the Model 2 R^2 from the Model 3 R^2. A nonzero answer is indicative of non-uniform DIF. Comparing the slopes of the two groups, Swaminathan and Rogers inferred non-uniform DIF if the difference in the slopes is not zero,

Table 3

Criteria for detecting DIF and Description of Item Characteristic Curves

	No DIF	Uniform DIF	Non-uniform DIF
Item characteristic curve Appearance	Curves overlap or are very close	Parallel curves but not coincident	Curves cross indicating different slopes; curves may have same intercept or different intercept
Zumbo (1999)	$b_2 = b_3 = 0$	$b_2 \neq 0$ & $b_3 = 0$	$b_3 \neq 0$
	$R^2_{\text{model 1}} - R^2_{\text{model 1}} = 0$	$R^2_{\text{model 2}} - R^2_{\text{model 1}} \neq 0$	$R^2_{\text{model 3}} - R^2_{\text{model 2}} \neq 0$
Swaminathan & Rogers (1990)	$b_1=b_2$ & $b_1=b_3$	$b_1 \neq b_2$ & $b_1 = b_2$	$b_1 \neq b_2$

meaning that the curves cross. Table 3 describes the nature of DIF with respect to item characteristics curves. See Appendix I for a summary of the LR equation variable meanings for applied DIF. LR simulation studies have also been conducted to compare DIF detection methods for uniform and non-uniform DIF for polytomously scored items. The review of these studies is beyond the scope of this research, however.

Effect size measures and LR. Though LR has many positive points, one drawback is its tendency to produce an overabundance of false positives or Type I error rates (Jodoin & Gierl, 2001). While it is desirable to identify items that contain DIF, it is highly undesirable to mark as compromised the test items that function as desired; this latter condition is precisely what happens when false positives occur. One reason for inflated Type I error rates is the increase in statistical significance which accompanies large sample size (Jodoin & Gierl, 2001). As sample size increases, power tends to increase, which thus leads to an increase in the number of DIF

Table 4

Classification of Negligible, Moderate, and Large DIF

Negligible DIF	Moderate DIF*	Large DIF*
0.13	0.13-0.26	> 0.26
Negligible DIF (A-level)	Moderate DIF** (B-level)	Large DIF** (C-level)
$R^2\Delta$-U < .035***	.035 ≤ $R^2\Delta$ - U < .070***	$R^2\Delta$ - U ≥ .070***

*Zumbo's (1996) suggestions based on Cohen's (1992, as cited in Jodoin & Gierl, 2001, p. 334)
**Must also have significant 2-df chi-square test to be flagged

items being identified. One solution for this problem is to introduce purification, which attempts to separate out a set of DIF-free items from the instrument being evaluated (French & Maller, 2007). Unfortunately, purification is statistically expensive, negating some of the positive points of LR. To counteract this problem, a measure of effect size can be used to indicate the magnitude of DIF. In this manner, test developers can make educated decisions about which DIF-containing items are the most problematic ones. It is rarely the correct decision to delete from a test all DIF-containing items; this strategy is simply too expensive and almost always unnecessary.

Guidelines for the classification of DIF are shown in Table 4. Zumbo (1999) and Gómez-Benito et al. (2009) set the goal of "empirically generating classification guidelines for negligible, moderate, and large DIF" (Jodoin & Gierl, 2001). Zumbo subtracted the model one likelihood model from the model three likelihood models to obtain the G2 statistic to measure effect size. According to Zumbo, "This [G2 statistic] modeling strategy is used to test whether the group and interaction variables are statistically significant over-and-above the matching criteria" (p. 27). Kanjee (2007) summarized Zumbo's work, as well as that of Nagelkerke (1991, as cited in Kanjee, 2007, p. 51), Zumbo and Thomas (1996, as cited in Kanjee, 2007, p. 51), and Jodoin and Gierl (2001), as follows:

> As in simple linear regression, it is possible to partition the R^2 statistic into components reflecting the effects unrelated to DIF (e.g., the τ_o and τ_1 parameters), those for uniform DIF (τ_2), and those for non-uniform DIF (τ_3). Thus, three values of R^2 are obtained. R^2_1 is derived from the model with only τ_0 and τ_1, R^2_2 is derived from the model that also includes τ_2, and R^2_3 is derived from the complete model that includes τ_3 as well. Using the notation of Jodoin and Gierl (2001), $R^2\Delta = R^2_3 - R^2_1$ reflects the overall DIF effect size, while $R^2\Delta - U = R^2_2 - R^2_1$ and $R^2\Delta - NU = R^2_3 - R^2_2$ reflect the effect size for uniform and non-uniform DIF respectively (Kanjee, 2007, p. 51).

Classification values modified using cubic regression was provided by Jodoin and Gierl. Meade (2010) provided a taxonomy of effect size measures, several of which are particularly applicable to the evaluation of non-uniform DIF.

Mantel-Haenszel Procedure Basics

The MH chi-square uses contingency tables to determine whether group membership and item performance are related. The big picture is that the odds of answering an item correctly is calculated for the reference and focal groups, and the performance on the test overall is taken into account. The MH test consists of a two-part calculation: the MH chi-square statistic which determines the presence or absence of DIF and whether DIF is uniform or non-uniform (Mantel & Haenszel, 1959), and the constant odds ratio for MH ($\widehat{\propto}$), which reveals the magnitude of the difference between the focal and reference groups. It is desirable to put the constant odds ratio for MH ($\widehat{\propto}$) on a different scale because the existing scale is asymmetrical with a lower bound of zero but an upper bound of infinity. It can be transformed to a log odds ratio (*beta*) and then the log odds ratio can be transformed to the ETS difficulty delta scale (*D*) by $D = -2.35 \ln (\alpha MH)$. If beta = 0 or D = 0, then the focal and reference groups performed the same on the item. If beta >

0 or D < 0, then the reference group is more likely to perform better and we say the reference group is favored. However, if beta < 0 or D > 0, then the focal group more likely performs better on the items and the focal group is the favored group. A confidence interval can also be calculated to estimate the range of *beta* or *D* (de Ayala, 2009).

The MH procedure used for many ETS programs focuses on statistical power (Dorans & Holland, 1993. After examinees are matched on an observed variable, such as total test score, MH uses an odds-ratio to compare the reference group to the focal group at each score level (Dorans, 1989). Reference and focal group data are organized using a two-by-two contingency table, which crosses group (focal and reference) with item performance (correct or incorrect response) for each ability level (Penfield, 2001). A tally is then collected for the focal group and the reference group to calculate the number of correct responses for each group on the item of interest. The likelihood of success is expressed as a ratio of the focal group to the reference group; this ratio is a measure of effect size. The significance test is distributed as a chi-squared statistic, which assesses the relationship between group membership (e.g., males versus females) and item performance across all ability levels (Penfield, 2001). Effective with samples sizes as small as 200, MH has one specific drawback. MH models were designed to find uniform DIF so their ability to detect non-uniform DIF is limited by design (Hambleton & Rogers, 1989).

Type I error inflation also results from repeated group comparisons using MH (Penfield, 2001). A suggested safeguard for dealing with the ineptitude of MH for detection of non-uniform DIF is to "routinely compare the direction of the difference in *p*-values for the two groups of interest across score groups and use graphing techniques" (Hambleton & Rogers, 1989, p. 333). Comparing the direction of the difference in *p*-values allows one to identify uniform and non-uniform DIF, respectively. If one group is favored through the entire range of test scores, then

uniform DIF occurs. If one group is favored over part of the range and the other group is favored over another segment of the range, then non-uniform DIF is indicated. Graphing techniques allow curves to be visualized; uniform DIF curves are parallel, shown in Figure 6, while non-uniform DIF curves, shown in Figure 7 cross at the point where the group that is favored changes.

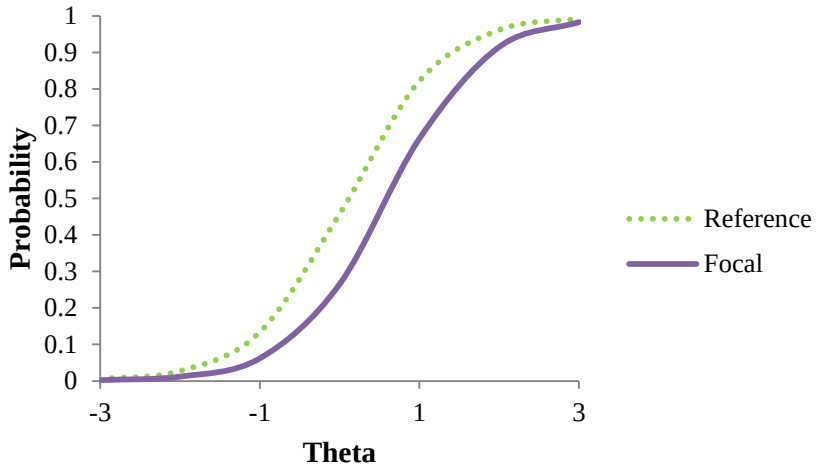


Figure 6. Parallel item characteristic curves illustrating uniform DIF.

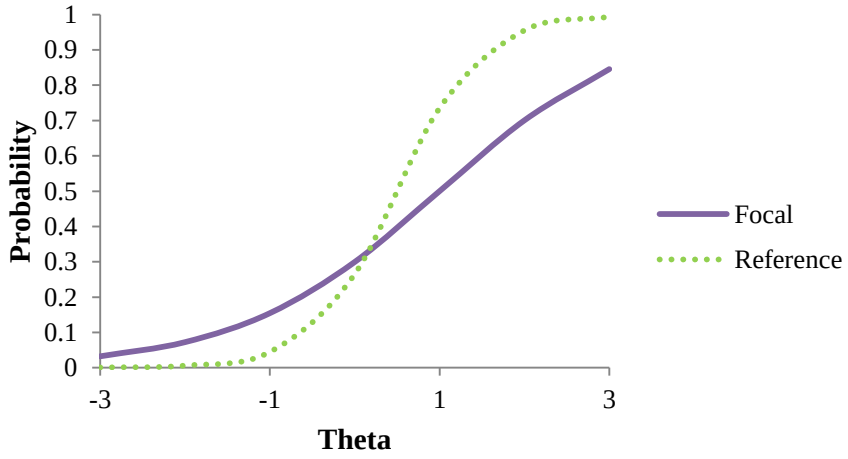


Figure 7. Item characteristics curves representing the focal and reference groups for a test item cross at the point where the favored group changes.

Two extensions of the MH procedure, the Mantel (Mantel, 1963) and the GMH (Mantel & Haenszel, 1959; Somes, 1986) can test for DIF at each score level (Thurman, 2009). Thurman (2009) reported, "The Mantel compares the item means after conditioning on a matching variable while the GMH compares the entire response distribution of the reference and focal groups" (p.18).

MH and detection of DIF. Though Zwick and Ercikan (1989) and Ackerman and Evans (1994) demonstrated the ability of MH to condition on more than one ability estimate, this matching strategy makes for foundationless and slow work**.** Since MH models lack latent variables, these models have no means to adjust for measurement error (Woods & Grimm, 2009). Woods and Grimm also suggested that MH methods are "are sensitive to differences in latent-variable variances between the focal and reference groups and that they lack robustness to non-normality despite being nonparametric procedures" (p. 340).

Summary of study results using MH procedures. Kwak, Nohoon, Davidson, & Davenport (1997) found the absolute mean deviation procedure outperformed the MH and unsigned MH in terms of power to detect non-uniform DIF as well as expected rates of false positives. After two iterations of purification, the MH procedures contained more false positives than they originally did. However, the purification process decreased the number of false positives for the absolute mean deviation procedure. Use of MH in current studies (Penfield, Alvarez, & Lee, 2001) has been limited to the detection of uniform DIF for which it was designed. Thurman (2009) recommended the GMH procedure over the ordinal logistic regression (OLR) and MH for use with polytomous data particularly when discrimination varies across items. Thurman examined DIF with respect to Type I error and power.

Hambleton and Rogers (1989) found the MH procedure easy to use and were impressed by its ability to handle smaller sample sizes, though they noted that since MH is analogous to a 1P IRT model it was not designed to detect non-uniform DIF. Indeed, they found in their 1989 study that MH did not detect non-uniform DIF. A modification to improve non-uniform DIF detection by the MH statistic proposed by Mazor, Clauser, and Hambleton (1994, as cited in Hidalgo & Lopez-Pina, 2004) split the sample of interest into high- and low-ability groups and implements the MH procedure on each group separately. When Hidalgo and Lopez-Pina (2004) used the modified MH procedure in their simulation study, they found that modified MH procedure for symmetrical non-uniform DIF detection rates approximated those of LR. For detection of asymmetrical non-uniform DIF, modified MH and LR performed similarly. However, LR was slightly superior to the modified MH procedure for the detection of symmetrical uniform DIF with correct identification rates of 68.75% and 61.25% respectively.

MH and detection of non-uniform DIF for polytomous data. Spray (1994) recommended logistic discrimination analysis over the MH procedure for the detection of non-uniform DIF in polytomous data as a result of the 1994 study comparing nominal and ordinal extensions of MH to logistic discrimination analysis for the detection of non-uniform DIF with large sample sizes ($N > 500$).

MH and missing data. Finch (2011) reported that MH procedures are robust to variety of types of missing data. Kwak et al. (1997) discovered that the absolute mean difference outperformed the MH and unsigned MH in terms of power to detect non-uniform DIF as well as expected rates of false positives. However, after two iterations of purification the MH procedures contained more false positives than they originally did. The purification process did decrease the number of false positives for absolute mean deviation.

Summary of LR, MH, and IRT Methods

IRT methods and the LR procedure perform best when detection of non-uniform DIF is a priority. Though the modified MH procedure performs similarly to LR on most levels, it is not as accurate at the detection of asymmetrical non-uniform DIF. Of all the methods, MH is the most economical and easy to understand, indicating that it might still be useful, especially if detection of asymmetrical non-uniform DIF were not an issue. Additionally, following the recommendations of Hambleton and Rogers (1989) concerning the safeguards to watch for the direction of difference in *p*-values and use of graphs could eliminate the need for use of sophisticated statistical methods to detect non-uniform DIF. For a comparison of the statistical and IRT methods for the detection of DIF see Appendix A.

CHAPTER 3: METHODOLOGY

As studies were evaluated for inclusion, it became evident that existing data for LR and MH presented in the same study would provide the most comparable data for comparing the two procedures. This study presented data based on background characteristics, such as type of study, and manipulated study characteristics, such as number of DIF items and sample size.

The study examined, through the lens of meta-analysis, the statistical methods LR and MH for evaluating the presence of and addressing DIF in testing instruments. One of the basic tenets of meta-analysis involves stating the problem at hand and outlining boundaries for the inclusion and exclusion of articles in a particular study (Curlette & Cannella, 1985). In this case, the problem at hand is the variety of different methods (e.g., LR and MH available to researchers wishing to identify DIF in testing instruments). Through meta-analysis, summary statistics such as effect size can be identified in primary research documents for both LR and MH and then compared on a common scale. Though the majority of included studies presented data as simply MH and LR, three studies used variations of either MH or LR or both. DeMars (2009) used 'standard' and 'nonlinear' LR and MH, Güler and Penfield (2009) used 'group' and 'interaction' for LR, and Li, Brooks and Johanson (2012) used 'raw' and 'deciles' matching for LR. For each study all types of LR and MH were averaged across all conditions of each study. Such a synthesis of the available body of literature provides DIF researchers quantitative information to use when analyzing test data.

Curlette and Cannella (1985) described five steps for conducting meta-analysis:

1. Define the problem and establish the criteria that will be used to determine admissible studies.
2. Search databases to locate studies for possible inclusion.

3. Determine and code the study characteristics.
4. Measure the study measures quantitatively on a common scale.
5. Aggregate the findings and relate those findings to the study characteristics.

In addition to the steps above, a *heterogeneity* test can indicate whether differences between studies are the results of true variations in the studies or if those differences can be attributed to sampling error, meaning chance. If heterogeneity tests do indicate differences between studies, those differences may be attributable to moderator variables (Sánchez-Meca & Marín-Martínez, 1997). In medical studies, heterogeneity is categorized as clinical, methodological, or statistical (Higgins & Green, 2008). According to Higgins (2008, p. 1158)

> It is generally accepted that meta-analyses should assess heterogeneity, which may be defined as the presence of variation in true effect sizes underlying the different studies. This assessment might be achieved by performing a statistical test for heterogeneity, by quantifying its magnitude, by quantifying its impact or by a combination of these.

In this study a statistical test for heterogeneity was conducted. The formula for the statistical test for heterogeneity is specific to the model used in the meta-analysis. If the effect sizes of individual studies do not possess one "true effect size," researchers can allow for heterogeneity through a random-effects model (Borenstein et al., 2009, p. 69). In the random effects model true differences between studies and therefore variations in effect sizes are anticipated. Examples of expected true differences in this study include incomplete disaggregation of data (DeMars, 2009; Narayanan & Swaminathan, 1996) and lack of comparable data for each condition of substantive study characteristics across studies. The computational formula for the heterogeneity test statistic (Q) in random-effects models is shown in equation 3.7. Conversely, if all studies are thought to share one effect size, and differences in effect size would be error-based, then a fixed-effect

model would be appropriate (Hedges & Vevea, 1998). Therefore, use of the fixed-effect model assumes that all studies share one true effect and that variations between studies are the result of sampling error or chance.

Borenstein et al. (2009) used statistical methods to allow for some differences between studies to occur as a result of random sampling variation instead of differences in effect sizes. They divide random error and real variance using formulas to quantify differences between studies and analyze them. For this study, Type I error effect size, provides the common scale for comparison of study characteristics across studies. Type I error effect size was calculated as a proportion of incorrectly identified DIF items versus total number of non-DIF items on simulated tests taken by simulated examinees. In addition to Type I error effect size three other effect size measures were used: Type. I error rates, deviation of Type I error rates from the .05 nominal level, and power rates. Another effect size measure for proportions is the arcsin transformed effect size, d_T (Gleser & Olkin, 2009), calculated by

$$d_T = 2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}. \quad (3.1)$$

Calculation of an additional effect size using this formula could potentially have produced a different value for Type I error effect size.

The simulation aspect of the study is important because the intentionally simulated DIF items allow for the calculation of a definitive Type I error because the DIF containing items are specifically constructed to contain DIF. In real data studies categorization of DIF items is affected by multiple variables such as, ability of examinees and the opportunity for examinees to belong to multiple groups, for example, high income and female, instead of a population being subdivided by dual group membership resulting in one focal and one reference group. In real data studies the presence of DIF is not a sufficient condition to remove a test item. An overall

view of the test is taken and generally only items exhibiting DIF in favor of one group or with large amounts of DIF magnitude, which is a measure of the amount of DIF an item contains, are removed from the test. Therefore, a simulated instrument with simulated examinees provides the opportunity to study Type I error rates in cases for which variables have been limited and the variations between studies have been documented. Most importantly because the DIF items have been specifically created in simulation studies, an exact number of DIF items is known which allows an exact Type I error rate to be calculated. The final coding sheet which includes summary effects, DIF detection method, test statistics and simulation study conditions can be found in Appendix J. Worked examples of Type I error effect size for MH and LR for included studies are shown in Appendix K. The preliminary DIF coding table is shown in Appendix L, and the preliminary data extraction worksheet headings are listed in Appendix M.

Literature Search

The section provides a description of how the uniqueness of the study was documented. Two web searches were conducted to search for existing literature which compared the ability of MH and LR to identify DIF items. A search of the ERIC at EBSCOhost database for *DIF and meta-analysis* yielded two articles: a DIF study summarizing 15 years of language testing (Ferne & Rupp, 2007) and a technical report providing a 2-year summary of research on the effects of testing accommodations (Thompson, Blount, & Thurlow, 2002). Changing the search string to *differential item functioning* and *meta-analysis* produced an additional two articles: one regarding psychometric approaches across independent studies (Bauer & Hussong, 2009), and another assessing DIF in writing assessments. Two searches of the Web of Science database for *DIF and meta-analysis* and *differential functioning and meta-analysis* yielded zero results. A Google search for *DIF* and *meta-analysis* unearthed an article using the outlier detection

approach with multiple groups using real data (Magis & De Boeck, 2012), as well as a mixture distribution conceptualization using real and simulated data (de Ayala, Kim, Stapleton, & Dayton, 2002). A summary by Hambleton, Clauser, Mazor, and Jones (1993) summarized six of their own studies completed over 12 years of research at the University of Massachusetts at Amherst pertaining specifically to IRT-based and MH DIF detection methods; four of these studies were simulation studies. Though much attention has been given to fairness and bias in testing, as well as DIF and statistical methods to identify DIF, a study has not been found which empirically summarizes the effectiveness of DIF detection methods using the format of a meta-analysis; this is the goal of the present study.

Real-data excluded studies. Examples of real-data studies exploring DIF detection methods include (a) a comparison of the Mantel-Haenszel (MH) and logistic regression (LR) procedures using chemistry and history test data from the College Board (Mazor, Kanjee, & Clauser, 1995); (b) delta plots with data from the helicopter aptitude test (Oosterhof, Atash,m & Lassiter, 1984); (c) Dorans and Kulick's (1986) summary of five studies that used the standardization approach on Scholastic Aptitude Test data; and (d) Wiberg's (2009) comparison of LR, MH, and log linear modeling with Swedish driving test data. One real data DIF meta-analysis, a math gender DIF study (Zhang, 2009), was located. An unpublished real-data dissertation examining gender and language DIF on students in Grades 3, 6, and 9 was excluded from my study (Zheng, Gierl, & Cui, 2007). See Appendixes N through R for a complete list of excluded studies organized by reason for exclusion.

Literature search process. The search for a meta-analysis on the subject of DIF began with a search of the ERIC at EBSCOhost and Education Full Text databases for *differential item functioning* and *meta-analysis*. The search returned six articles. One of these primary research

studies was a duplicate. The remaining five articles addressed DIF and meta-analysis in the following ways: (a) the use of integrative data analysis to summarize data from two longitudinal studies with data about alcohol use (Bauer & Hussong, 2009); (b) assessing DIF in writing assessments using the GMH statistic and logistic discriminant function analysis with a meta-analysis based method on actual data from eighth-grade students (Welch & Miller, 1995); (c) searching for a gender-by-item interaction among males and females on multiple-choice math items (Bielinski & Davison, 1998); (d) categorically summarizing testing data for students with disabilities by examining how accommodations affected test score by tallying data and placing it into categories (Thompson, Bount & Thurlow, 2002); and (e) qualitatively reviewing 27 studies to summarize five sets of characteristics important to language testing (Ferne & Rupp, 2007).

As far as I have been able to determine by reviewing the literature, the current study is unique. Having completed the first step by documenting the uniqueness of my study, the second step of the literature search was initiated. The aim of this aspect of the study was to systematically obtain papers, published and unpublished, pertaining to the use of LR and MH for the evaluation of DIF for the time period spanning 1975 to 2013. Because effect size is the gold standard for comparing studies in meta-analysis (Borenstein et al., 2009), the initial search in the ERIC at EBSCOhost database was done for *DIF* and *effect size*. This search yielded 62 articles. These articles were then screened by hand to find articles that contained *DIF effect size*. Approximately 23 articles were found to have those key search terms. Next, the methods section of each article was studied to determine whether *DIF effect size* would be an appropriate outcome variable for meta-analysis. After careful study of the articles, "DIF effect size" was found to be unsuitable as an outcome variable for meta-analysis, because it measured the amount of DIF, also referred to as DIF magnitude, present in each DIF item. In meta-analysis the term

effect size refers to a summary statistic used to compare overall statistical differences in studies. The initial search for 'DIF and effect size' intended to find any existing studies comparing DIF across studies utilizing the meta-analytic summary statistic referred to as effect size. This 'DIF effect size' or 'DIF magnitude' was not an appropriate outcome measure for this study since it measured the amount of DIF present in particular test items and was therefore unrelated to type I Error. At this point, it was clear that a new outcome variable was needed. Type I Error effect size was chosen as the outcome variable and was calculated by dividing the number of incorrectly identified DIF items (e.g. false positives) by the total number of items on each test. Additional effect size measures utilized in the meta-analysis were Type I error rates, deviation of Type I error rates from the nominal .05 level, and power rates. In the paragraphs that follow the term effect size is discussed as it pertains to the meta-analytic methods employed in this study. The outcome variable for the study which was used to conduct a statistical comparison of included studies is referred to as type I Error effect size.

In this study effect size is calculated by taking the difference of Type I error effect size for LR and MH, two proportions; it is then divided by the pooled standard deviation. These calculations results in studies being placed on a common scale, which is necessary for meaningful comparison. Borenstein et al. (2009, p. 18) presented four considerations that should be taken into account when searching for an effect size:

- Effect sizes from different studies can be compared.
- Estimates of effect size can be computed from data published in studies.
- The sampling distribution of the effect size known as the confidence interval can be calculated.
- The effect size is presented as an interpretable metric for researchers in the field.

These principles were used to begin the search for a way to measure the effectiveness of DIF detection across studies including results that could also be converted to an effect size. Type I error was mentioned frequently, and many studies contained Type I error data. Because Type I error is the proportion of false positives (e.g., the number of items identified as DIF which were indeed unbiased items), it was selected as the outcome variable for the meta-analysis. With the selection of this variable came the idea to narrow the inclusion criteria so only simulation studies would be included; this change worked well with Type I error, because simulated DIF items are created using exact parameters, and the exact number of true DIF items is known.

Since DIF effect size was not an appropriate outcome variable, the second step was to use the "find" function in Adobe Acrobat to search each possible article for Type I error data. Since Cochrane guidelines recommend searching more than one data base (Higgins & Green, 2008), a subsequent search was conducting using the Web of Science database. A search for *differential item functioning* produced 1,613 results. The results were refined by specifying inclusion of the following terms: simulation study, Mantel-Haenzsel, logistic regression and Type I error. Of the resulting 17 studies, four had already been marked for inclusion (Güler & Penfield, 2009; Kim & Oshima, 2012; Li, Brooks, & Johanson, 2012; Vaughn & Wang, 2010). Hand screening of the reference lists of the remaining 13 articles yielded two additional articles containing potentially usable data. Of those two, one was only available only in Spanish (Raver, Aliste & Muniz, 2000); the other article was marked for inclusion. Appendix S depicts the search process.

Research Design

Limitations of Meta-Analysis

General drawbacks of meta-analysis include (a) difficulty in obtaining necessary data, (b) tendency to use published research that suffers from publication bias, and (c) appropriate use

of inclusion and exclusion criteria. Inclusion of too many studies makes it difficult to summarize results accurately. Failure to include an adequate number of studies can make it difficult to paint a true picture of the data. Conducting meta-analysis places one at the mercy of other researchers. If the papers being synthesized do not represent rigorous research, neither will the meta-analysis.

Confounding occurs if a variable with the potential to change the effect size or variation between studies is not measured or included with the study results (Littell et al., 2008). Confounding could be an issue with the current meta-analysis due to the small number of studies that met inclusion criteria. Confounding is a problem because study characteristics that might have been pooled to tease out their effect on the independent variable will be either ignored or averaged due to the lack of a method to include them in a meaningful manner. Excluded studies are listed in Appendixes N through R. They are organized by those containing: (a) LR and MH data not in usable form; (b) either LR or MH data, (c) neither LR nor MH data; (d) solely real data; and (e) Type I error data. Only studies containing MH, LR, and Type I error data were admissible; included studies are denoted with an asterisk in the References section. Included studies with data type and location are listed in Appendix G.

Inclusion and Exclusion Criteria of Studies

Ten of the 62 screened articles met inclusion criteria for the meta-analysis. The inclusion criteria specified that the study (a) used LR and MH to detect differentially functioning items, (b) was a simulation study, (c) contained Type I error data either in summary form or by condition, and (d) was published between 1975 and 2013. Studies were excluded if they (a) only used real data; (b) did not contain Type I error data, including studies containing data (e.g., means and standard deviation) that may have been converted to effect size data; (c) examined either LR or

MH, but not both; (d) presented Type I error results, but not the raw Type I error data needed to calculate the Type I error effect size; or (e) were not available in English.

Coding

In this dissertation data extraction was conducted independently. Since Type I error effect size formulas were not included in any of the software packages, a researcher created Microsoft Excel spreadsheet was used for coding. Initially, every possible study characteristic (Appendix L) was coded and several iterations of the coding worksheet were adapted. Coding worksheets were based on Thurman's (2009) dissertation, which provided a template for organizing the coding process (see Appendix J). Thurman's work was helpful because it shared most of the study characteristics of the articles included in the current meta-analysis and gave attention to data generation procedures, allowing methodological, substantive, and background study characteristics to be examined.

Background characteristics are unchanging aspects of a study. Examples of background characteristics (Curlette & Canella, 1985), also called fixed study parameters, for the included studies are author(s), publication date, and type of study (e.g., simulation study). Each of the studies of the current meta-analysis included additional unchanging study characteristics, but these were not uniform across the studies. Thus, they are treated as substantive study characteristics with respect to data analysis.

Substantive study characteristics (Curlette & Cannella, 1985) have the potential to affect the outcome variable. The final code sheet is organized by substantive study characteristics, found in Appendixes T through X, and methodological study characteristics, found in Appendixes B through E. These appendices list characteristics shared by most of the studies included in the meta-analysis. Methodological study characteristics are tied to the steps carried

out during primary research (Curlette & Canella, 1985). In an ideal situation, methodological and substantive study characteristics would be clearly delineated. However, the small pool of studies in the current meta-analysis and the use of simulation studies blur this line. Creation of subgroups is problematic because relatively few studies were included in the meta-analysis. Under these conditions, slight variations with respect to generating models and differences in item parameters, which could be considered methodological, actually had the potential to affect the outcome variable.

Use of Effect Size to Compare Studies

According to Borenstein et al. (2009), *effect size* is an appropriate term to describe an index used to quantify the difference between two groups or variables. Simulation studies that generate DIF items and then use LR and MH methods to identify DIF items gauge success by the rate of Type I error. This Type I error rate reveals the proportion or percentage of flagged DIF items that are actually DIF-free. Ten studies met inclusion criteria for the present research. Seven studies were marked for exclusion (Appendix N) because the data LR and MH they displayed is not in usable form for this meta-analysis: means and standard deviations (Chan, 2000, pp. 183,185; Hidalgo & Lopez-Pina, 2004, p. 912; Kim et. al, 2007, pp. 101, 105, 111); correlations and bias statistics (Hambleton & Rogers, 1989, p. 326); absolute bias (Woods & Grimm, 2011); overall bias (Robitzsch & Rupp, 2009, p. 28); and *p*-value and standard error (Wiberg, 2009, p. 50). Though data from the DeMars (2009) study was not originally in a usable form, emailing the author resulted in receipt of a detailed Microsoft Excel spreadsheet containing usable data. Correspondence with authors of other studies was not fruitful.

The following 16 studies, shown in Appendix O, were excluded even though they contained Type I error data because they did not contain both LR and MH methods for DIF

analysis: Fidalgo et al.'s (2007, p. 305) MH and Loss Function; Finch and French's (2008, p. 751) SIBTEST, IRT likelihood ratio and LR; Jodoin and Gierl's (2001, p. 341) SIBTEST and LR; Kanjee's (2007, p. 56) two variations of LR were applied to uniform and non-uniform DIF populations; Penfield's (2001, p. 244) three variations of MH; Gómez-Benito et al.'s (2009, p. 29) MH and SIBTEST; Hidalgo and Gomez's (2006, p. 819) Multinomial Logistic Regression and Discriminant Logistic Analysis with and without purification; Spray and Miller's (1994, p. 13) MH and logistic discriminant function analysis; Su and Wang's (2005, pp. 328-333) variations of MH, average signed area, logistic discriminant function analysis and partial credit model; Wang and Su's (2004, pp. 131-137) variations of MH; and Zwick, Thayer, and Mazzeo's (1997, p. 335) standardized mean difference, Mantel and SIBTEST. Excluded studies organized by reason for exclusion are shown in appendixes N through R. Page numbers indicate the location of Type I error in each document.

Model Selection and Calculations

Fixed-effect versus random-effects. Meta-analysis can be conducted under one of two models: fixed-effect or the random-effects model. The fixed effect model assumes that one true effect, which is unchanging, exists for all studies save the occurrence of sampling error. This means that factors influencing the effect size do not vary from study to study. In the random-effects model the true effect is free to vary with each study (Borenstein et al., 2009). This means that differences between effect sizes of studies in the fixed-effect model are treated as sampling error and not calculated; while the effect sizes of random-effects models are expect to vary due to differences in the studies. This meta-analysis uses a random-effects model since DIF percentage, test length and replications are substantive study characteristics which do not have equal counterparts across included studies. Statistical formulas for the two models are shown in

this chapter. In the fixed-effect model only one variation is calculated, the variation within studies. For random-effects models the variation between studies is calculated, as well as a value for total variance which is the sum of differences in effect sizes, represented using standard deviations, between and within studies.

Influence of substantive study characteristics on model selection. The included studies in the current meta-analysis share four substantive study characteristics (impact, sample size, DIF percentage, and test length) in addition to specified inclusion criteria. Test length is a variable for three studies (DeMars, 2009; Kim & Oshima, 2012; Swaminathan & Rogers, 1990), yet is fixed for the other studies while still varying in length from 20 to 100 items. Sample size provides another instance of differing similarities with equal groups of 250 and 500 as popular sizes (Rogers & Swaminathan, 1993; Swaminathan & Rogers 1990; Vaughn & Wang, 2010). Among these commonalities, the additional conditions of equal ability distribution (impact = 0) for eight included studies (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1993; Rogers & Swaminathan, 1993; Swaminathan & Rogers 1990; Vaughn & Wang, 2010) versus unequal ability distribution (impact = 1) for six included studies (de Ayala, 2002; DeMars, 2009; Güler & Penfield, 2009; Li et al., 2012, Narayanan & Swaminathan, 1993; Vaughn & Wang, 2010) adds an additional layer of analysis. DIF percentage is treated as a variable by three studies (de Ayala et al., 2002; DeMars, 2009; Narayanan & Swaminathan, 1996) using 0%, 10%, 20%, and 30%, for the rest of the studies, DIF percentage is fixed at 0%, 10%, 12%, 15% or 20%.

Influence of methodological study characteristics on model selection. The trend of similar differences continues with the data generation facet of the studies. Although most studies used an IRT 3PL to generate data, two (de Ayala et al., 2002; Li et al., 2012) opted for the 2PL,

while Rogers and Swaminathan (1993) used the 2PL to indicate *good fit* and the 3PL to indicate *poor fit.* The number of DIF items varies, with four studies (de Ayala et al., 2002; DeMars, 2009; Güler & Penfield, 2009; Kim & Oshima, 2012) using the same number (6). The most common number of replications (100) was used by five studies (DeMars, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). The lowest number of replications is 20 (Swaminathan & Rogers, 1990), while the highest is 10,000 (Li et al., 2012). Due to these methodological differences, the random-effects model was chosen to classify the studies in the current meta-analysis.

Calculations for the random-effects model. Several differences in calculations occur as a result of choosing a random-effects model. Because a random-effects model was chosen for the current meta-analysis, the following formulas were used to conduct the calculations. The first difference is weighting of studies. In a fixed-effect model, the weighting can be accomplished using sample size because the true effect is assumed to be the same. In a random-effects model, however, study weights are more balanced with larger studies receiving less weight and smaller studies receiving more weight than they would under a fixed-effect model (see equation 3.8; Borenstein et al., 2009). In this meta-analysis weighting conventions have been altered so that the weight is determined by the number of replications carried out for each study instead of the study size.

In the random-effects model, researchers assume there are "real differences between all the studies in the magnitude of the effect" (Borenstein et. al, 2009, p. 61). From study to study, the *random effect* is the standard deviation, which represents the variation in the true magnitude. For a random-effects study, the following calculations are necessary: effect size (here, Type I error effect size), variance within (V_y), variance between (T^2), total variance ($V_y + T^2$), weight

(W*), weight times effect size (W*Y), and summary effect (M*). An asterisk is used to distinguish random-effects formulas from those used with fixed-effect models. Random-effects formulas are as follows:

Type I error is calculated by

$$\frac{\text{number of incorrectly identified non−DIF items}}{\text{total number of non−DIF items}} = \text{Type I error}, \qquad (3.2)$$

and Type I error, for which data is available for each included study, allows LR and MH methods of DIF detection to be compared across studies, thus serving as the effect size for the meta-analysis,

$$\text{Effect Size} = \text{Type I error}. \qquad (3.3)$$

The variance within, represented by $V_{y,}$

$$V_{yi} = sd^2, \qquad (3.4)$$

estimates the differences existing within a particular meta-analysis by squaring the standard deviation for each study. This value will be calculated for each of the ten studies.

The variance between, represented by τ^2,

$$\tau^2 = \frac{Q-df}{C} \quad \text{if } Q > df, \qquad (3.5)$$

or

$$\tau^2 = 0 \quad \text{if } Q < df, \qquad (3.6)$$

"is defined as the variance of true effect sizes" (Borenstein et al., 2009, p. 114). The variance between estimates the differences studies using a series of three calculations involving: the degrees of freedom, represented by *df*,

$$df = k - 1, \qquad (3.7)$$

where the number of studies is represented by k,

$$k = \text{the number of studies}, \qquad (3.8)$$

the Q-Statistic, represented by Q,

$$Q = \sum_{i=1}^{k} W_i^* \, Y_i^* - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}, \quad (3.9)$$

where the study number is represented by i,

$$i = \text{study number}, \quad (3.10)$$

and the weight is represented by W*,

$$W^* = W_i^* = \frac{1}{V_i^*} = \frac{1}{V_{yi} + \tau^2}, \quad (3.11)$$

where the total variance, represented by V_i^*,

$$V_i^* = V_{y_i} + T^2, \quad (3.12)$$

is the sum of the within-study variance which is the variance in the effect size for the particular study, represented by V_{y_i} ,

$$V_{y_i} = \text{variance within}, \quad (3.13)$$

added to the between-study variance, represented by τ^2 which is estimated from the observed effects, (3.4). If the value of the Q-statistic is greater than the degrees of freedom, a value is needed to put the variance between, τ^2, "back into its original metric and also make it an average" (Borenstein et. al., 2009, p .114). The quantity represented C,

$$C = \sum W_i^* - \frac{\sum W_i^{*2}}{\sum W_i^*}, \quad (3.14)$$

is used to accomplish that (Borenstein et al., 2009, pp. 109, 114-115).

Two Type I error effect sizes were calculated. Rosenthal (1994, p. 237) provides formulas for the calculation of *d'* (3.14). The *d'* calculation is performed by subtracting the first proportion from the second,

$$d' = (p_1 - p_2). \quad (3.15)$$

Additional formulas needed for the analysis include the weighted mean or summary effect,

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}, \quad (3.16)$$

The reciprocal of the sum of the weights is estimated as the variance of the summary effect,

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*}, \quad (3.17)$$

And the square root of the summary effect is the estimated standard error of the summary effect,

$$SE_{M^*} = \sqrt{V_{M^*}}. \quad (3.18)$$

Additionally, the 95% lower,

$$LL_{M^*} = M^* - 1.96 \times SE_{M^*}, \quad (3.19)$$

and 95% upper limits,

$$UL_{M^*} = M^* + 1.96 \times SE_{M^*}, \quad (3.20)$$

for the summary effect are computed (Borenstein et al., 2009, pp. 73-74).

After performing the necessary summary effect calculations, the next step was to make sense of the variations in the effect size. Borenstein et al. (2009) referred to observed differences in effect size as heterogeneity of effect sizes. These differences include not only true variations, but also random error. The spurious nature of the values necessitated the use of a series of formulas to address questions about the variation. The statistics include the Q statistic, "the results of a statistical test based on the Q statistic (*p*), the between-studies variance (T^2), the between studies standard deviation (T),

$$\sqrt{T^2}, \quad (3.21)$$

and the ratio of true heterogeneity to observed variation (I^2)"

$$I^2 = \left(\frac{Q-df}{Q}\right) x\ 100\%, \quad (3.22)$$

(Borenstein, et al., 2009, p. 105). Because the summary effect (M*) has already been calculated, formula 3.15 (Borenstein et al., 2009, p. 109) can be used to calculate Q, which is the weighted sum of squares,

$$Q = \sum_{i=1}^{k} W_i^* (Y_i^* - M^*)^2. \quad (3.23)$$

In this study MetaAnalyst software was used to carry out the calculations for the random-effects model.

Independent and Dependent Variables in Meta-Analysis

The dependent, or outcome, variable in this study is Type I error effect size. Type I error inflation has been associated with higher item discrimination values, or *a* parameter values, (DeMars, 2009) and unequal ability distributions (Narayanan & Swaminathan, 1996; Penfield, 2001). Study characteristics, the independent variables in the present meta-analysis, are variables present across studies that could contribute to differences in the Type I error effect size between studies. Curlette and Cannella (1985) divided study characteristics into two groups: substantive and methodological. Substantive study characteristics are those that can influence the outcome variable of the study, while methodological characteristics are specific to the steps taken to carry out the study. Background study characteristics refer to unchanging aspects of a study such as the author, title, or type of study. In the preliminary phases of the study, possible study characteristics included statistical and IRT methods used for DIF detection, use of ANOVA to assess effects, use of effect size for classification purposes, ability difference (equal or unequal), sample size, DIF percentage, DIF magnitude, power data, and specifics of data generation. These study characteristics are represented in Appendixes B through E. Several suitable substantive study characteristics emerged during the coding process: sample size, ability differences (impact), DIF percentage, and test length (DIF items plus non-DIF items); these characteristics

are reflected in Appendixes V through X. The following study characteristics were categorized as methodological since they were used to create the unique items and examinees for each simulation study: number of studied items (DIF-containing items), *a* parameter values (discrimination), *b* parameter values (difficulty), DIF magnitude (change in *b* parameters between the focal and reference groups),and nature of DIF (uniform or non-uniform). Methodological study characteristics are summarized in Appendixes B through E.

Substantive study characteristics. Substantive study characteristics can influence the outcome variable or dependent variable of the study, which in this case is Type I error. Because inclusion criteria for this meta-analysis specifies simulation studies, one way to separate substantive study characteristics from methodological ones is to assign methodological status to study characteristics pertaining to simulation of data and substantive status to those characteristics used as a variable in any study.

Ability distribution differences. Though group ability difference did not vary for all included studies, Type I error tends to be inflated when groups exhibit ability difference (Narayanan & Swaminathan, 1996). Group ability difference, or impact, was a variable for the following studies: DeMars (2009), Güler and Penfield (2009), Li Brooks, and Johanson (2012), Narayanan and Swaminathan (1996) and Vaughn and Wang (2010). Ability distribution was varied by making changes to the normal distribution (N[0,1]), mean = 0, SD =1, shown in Appendix T.

Sample size. According to Kim (2010), sample size is a key variable for DIF detection. A summary of sample size and replication information for the studies can be found in Appendixes X and E, respectively. Though the danger with small sample sizes is the sin of omission when searching for DIF items, with larger sample sizes the keen accuracy produced could actually

point out DIF-containing items with so small a DIF magnitude that discovery of these items would be irrelevant. De Ayala et al. (2002) also used a wide spread between focal and reference groups (500/2,500) based on the work of Zwick, Donoghue, and Grima (1993). In addition, Güler and Penfield (2009) provided an adequate berth to focal and reference groups with regard to sample size (300/1,000), stating the common use of minimum sample sizes of 200 to 250 (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) to provide adequate power for LR and MH.

DIF percentage. For three included studies, the 10% DIF condition was not specified. Güler and Penfield (2009), as well as DeMars (2009), included this condition as a constant. Two studies simulated a DIF percentage of zero (Li, Brooks, & Johanson, 2012; Rogers & Swaminathan, 1993). Though 2002 study of de Ayala et al. as well as DeMars (2009) and Narayanan & Swaminathan (1996) treated DIF percentage as a variable, on de Ayala et al. presented disaggregated results ideal for comparison. Swaminathan and Rogers (1990) opted for 20% DIF and subsequently Rogers and Swaminathan (1993) tried a 12.5% DIF condition. Specific conditions of DIF percentage and test length for each study are summarized in Appendixes U and W.

Test length. Studies highlighted a variety of test lengths for differing reasons, and the rationale for using various test lengths were not given by all. Rogers and Swaminathan (1993) favored a 40-item test because of its relatedness to the lengths of subsets on standardized tests. Subsequent studies followed their lead (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1994). Güler and Penfield (2009) opted for a 60-item test due to its similarity to standardized tests. Swaminathan and Rogers (1990) maintained that the longer the test, the more accurate a measure of total score produced. Test length factors into the calculation of LR and MH. For LR,

total score serves as the predictor in the model, while for MH total score is used as the criterion for grouping test candidates. Having previously used tests with lengths of 40 and 60 items, they added an 80-item test to experiment with their assertions of the influence on test scores on LR and MH.

Methodological study characteristics. Methodological study characteristics pertain to the structure guiding the steps of the study. Study characteristics governing data generation could have been considered substantive for this meta-analysis. Due to their possible effect on the outcome variable, Type I error, variables pertaining to data simulation have been considered carefully in this meta-analysis. These include the 2PL or 3PL generating model and method for generating item parameters, which are shown in Appendix B.

Difficulty and Discrimination. Studies used either 2PL or 3PL IRT models to simulate data.. Therefore, it was straightforward to introduce DIF through manipulation of the discrimination (*a* parameter) or difficulty (*b* parameter). Changes made to the *a* parameter brought about non-uniform DIF, while changes to the *b* parameter resulted in uniform DIF. Most studies varied the *b* parameter as a condition or to create DIF. Appendix D contains a basis for comparison of the difficulty and discrimination parameters across studies. Parameter values are represented either as a change in the parameter, the actual parameter values or both depending on reporting methods of the study. DeMars (2009) asserted that Type I error inflation increases with increasing item discrimination (*a* parameter). DeMars' study manipulated item discrimination using five levels of the *a* parameter beginning with 1.2 and increasing in increments of 0.2 through 2.0. Several other studies varied item discrimination (Güler & Penfield, 2009; Li, Brooks, & Johanson (2012); Rogers & Swaminathan, 1993).

DIF magnitude and nature of DIF. This substantive study characteristic, in fact, pertains to the amount of DIF present in a test item. Herein lies part of the appeal of simulation studies, which allow researchers to create data sets showcasing exactly the variable characteristics they aim to study. Swaminathan and Rogers (1990) used "nature" to group uniform and non-uniform DIF, shown in Appendix C. Uniform DIF is simulated by altering the *b* parameter of the item. If the *a* and *b* parameters were the same for focal and reference SD_r = standard deviation (reference group), SD_f = standard deviation (focal group) groups, the item curves would be superimposed illustrating a situation of no DIF. Figures 1 and 2 illustrate this phenomenon for the 1PL and 2PL models, while Figure 3 shows the 3PL no DIF situation in which the *c* parameter would also be held constant. For included studies, the *c* parameter value was held constant at 0.2, which is considered to be a typical guessing parameter for multiple-choice questions having five answer choices (Hambleton et al., 1991). Differences in *b* parameters refer to the differences between the difficulty, or item location, values of the reference group versus the focus group. Increasing the *b* parameter moves the item curve so that it is higher on the graph as the curve moves to the right, shown in Figure 4, while keeping its slope (*a* parameter, or discrimination) constant, thus creating a test item that is more difficult for all members of the focal group. Typically, the focus group is often the smaller group, and is considered to be the group experiencing DIF. The focus group, then, is often the group for which the item is more difficult. "Moderate" and "medium" DIF magnitude values are synonymous with changes in the *b* parameter of 0.25 to 0.5, while "high" DIF magnitude syncs up with changes in the *b* parameter of 0.64 to 1.0. This information is presented in Appendix C; since verbiage in Appendix C matches that of the source articles, terms may vary.

Non-uniform DIF occurs when *a* parameter values change, resulting in different slopes for the focal and reference groups. In Figure 5, the dotted line shows a group that has more success answering easier questions, e.g., those in the range from -2 to 0 on ability or theta level, while the same group finds questions in theta range 0 to 2 more difficult. The solid line, representing the reference group, represents a population doing poorly on easier questions, e.g., theta range -2 to 0, yet having greater success with more difficult questions, e.g., theta range 0 to 2. Such a scenario could be explained by a more skilled population, the reference group, making careless errors on easier questions while focusing more intently on more difficult questions, while a second population, the focal group, is able to answer easier questions correctly, but does not find success with more difficult questions.

Replications. The number of replications of included studies varied from as few as 20 (Swaminathan & Rogers, 1990) to as many as 10,000 for Li, Brooks, & Johanson (2012). Rogers and Swaminathan (1993) used 100 replications for their research, and Kim (2010) cited the National Council on Measurement in Education's statement that 100 replications was the common choice. DeMars (2009) varied the number of replications from 100 to 300 with three different test lengths.

Summary

The research pursued in the current dissertation involved conducting a meta-analysis of simulation studies that explored the effects of using MH and LR to identify DIF on the Type I error rate for correct identification of differentially functioning items. The techniques of Borenstein et al., (2009) for implementing meta-analysis are thorough, inclusive, and statistically sound, and it is on this foundation the methodological approach of this dissertation is laid. Some

issues raised by Glass are still relevant, but the calculations are based on Borenstein et al.'s formulas.

CHAPTER 4: RESULTS

The purpose of this study was to compare the performance of LR and MH statistical methods for DIF detection under various simulated conditions via meta-analysis. The following research questions directed the study:

1. Under various conditions in Monte Carlo computer simulations, how do the Type I error rates compare for LR and MH?
2. How does each LR Type I error proportion & MH Type I error proportion compare to the accepted detection rate of .05?
3. How do the following substantive study characteristics affect Type I error effect size: impact, sample size, DIF percentage, and test length?
4. How do Type II error rates compare for those studies displaying power data?

Answers to each of the research questions will be presented in this section. Ten articles met inclusion criteria while 57 articles were excluded from the meta-analysis (Appendixes N-S). In meta-analysis effect size is the gold standard for comparing studies (Borenstein et al., 2009). Type I error evaluated at the .05 level for LR and MH was needed to calculate Type I error effect size, which was used to compare the number of false positives incurred for each included study.

Research Question 1: Comparison of Type I Error Rate by Condition and MH and LR

Seven studies provided Type I error rates. For the remaining three studies, Type I error rate was calculated from the data provided. Table 5 lists each included study and the manner in which Type I error rate was obtained. De Ayala et al. (2002) provided the number of times each unbiased item was identified as biased, so that number was divided by the number of replications, 50, to calculate Type I error rate. Since Narayanan and Swaminathan (1996) and

Table 5

Type I Error Summary: Data Location by Study and Sample Calculations

	Type I Error				
Study	**Rate**	**Raw Data**	**Data Format**	**Sample Calculation**	**Unbiased Items (replications)**
de Ayala el al. (2002)	p. 255, 257, 259, 261, 263	p. 254-263	Number of Times Identified as DIF	2/50 = 0.04	26-30 (l50)
DeMars (2009)	excel spreadsheet from author	na	Type I error	na*	16-56 (100-300)
Güler & Penfield (2009)	p. 323	na	na	na	1 (200)
Herrera & Gomez (2008)	p.748	na	Rate of False Positives	na	88 (100)
Kim & Oshima (2012)	p.. 465-466	na	Type I error Rate	na	17-37 (100)
Li, Brooks & Johanson (2012)	p.854, 857, 858	na	Type I error Rate	na	50 (10,000)
Narayanan & Swaminathan (1996)	p. 267	na	Type I error Rate Percentage	4.4/100 = 0.04	24 (100)
Rogers & Swaminathan (1993)	na	p. 112	Number Unbiased Items falsely identified	6/100 = 0.06	35 (100)
Swaminathan & Rogers (1990)	na	p. 367	Number of Items Flagged as Biased	1/32 = 0.03 1/48 = 0.02 3/48 = 0.0625 1/64 = 0.0156	32 (20) 48 (20) 48 (20) 64 (20)
Vaughn & Wang (2010)	p. 948	na	Type I error Rate	na	32 (10,000)

*Type I error rate provided in article

Rogers and Swaminathan (1993) each performed 100 replications in their studies, dividing the number of unbiased items mistakenly identified as DIF-containing by 100 resulted in the Type I error rate for those studies. Swaminathan and Rogers (1990) provided 'Number of Items Flagged as Biased' over 20 replications, therefore dividing the number of false positives for each condition by the number of unbiased items provided the Type I error rate for their study. The number of unbiased items was calculated by subtracting the number of DIF items from the total number of items for each condition.

Presentation of Type I Error Data

Type I error rates were presented in two different ways. Of the ten included studies, eight presented Type I error data by condition (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) and two studies presented Type I error data by item (de Ayala et al., 2002; Li, Brooks & Johanson, 2012). Assimilation of Type I error data by condition required selecting the most relevant data for comparison from each study. Data were segregated by impact and sample size conditions, but data were averaged across some conditions of test length and percentage of DIF. Narayanan and Swaminathan (1996) provided Type I error percentages for sample size, impact and DIF percentage in aggregate form only. For this reason data from that study are recorded twice in the graphing spreadsheets: once for sample size and a second time as an average for impact. A discussion of the manner in which data were assimilated for graphical presentation by Type I error follows.

Impact and Sample Size

Each included study contained data either for the condition of impact equals zero (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012;

Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) or impact equals one (de Ayala et al., 2002; DeMars, 2009; Güler & Penfield, 2009; Li, Brooks & Johanson, 2012; Narayanan & Swaminathan, 1996).

Sample size was grouped in conjunction with either equal ability (i.e., impact = 0) or unequal ability differences (i.e., impact = 1). Additionally, sample size was categorized as equal with the focal and reference groups containing the same number of examinees or unequal with the focal and references groups containing different numbers of examinees, generally with the focal group being the smaller group. Equal sample size groups were classified as small (200-300), medium (500-700) or large (1,000-2,500). Boundaries for the groups were delineated based on available data contained in the included studies as well as naturally occurring breaks. Categories were selected to increase the ability to discuss research on what we see as small, medium and large and to clearly define groups in the manner of extreme group studies. These steps to increase clarity were taken realizing that these ranges do not allow for the incorporation of all effect sizes. While use of these boundaries resulted in elimination of a few specific conditions from a small number of studies, no studies were excluded based on these boundaries. Five studies contained small, equal sample size data (DeMars, 2009; Güler & Penfield, 2009; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang; 2010). Medium, equal sample size data existed for six studies (Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010), and five studies contained large, equal sample size data (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Li, Brooks & Johanson, 2012; Vaughn & Wang, 2010). Type I error data for MH and LR for impact equal to 0 with

Numbers atop the bars are the Type I error rates. All studies had nominal Type I error rate of .05.

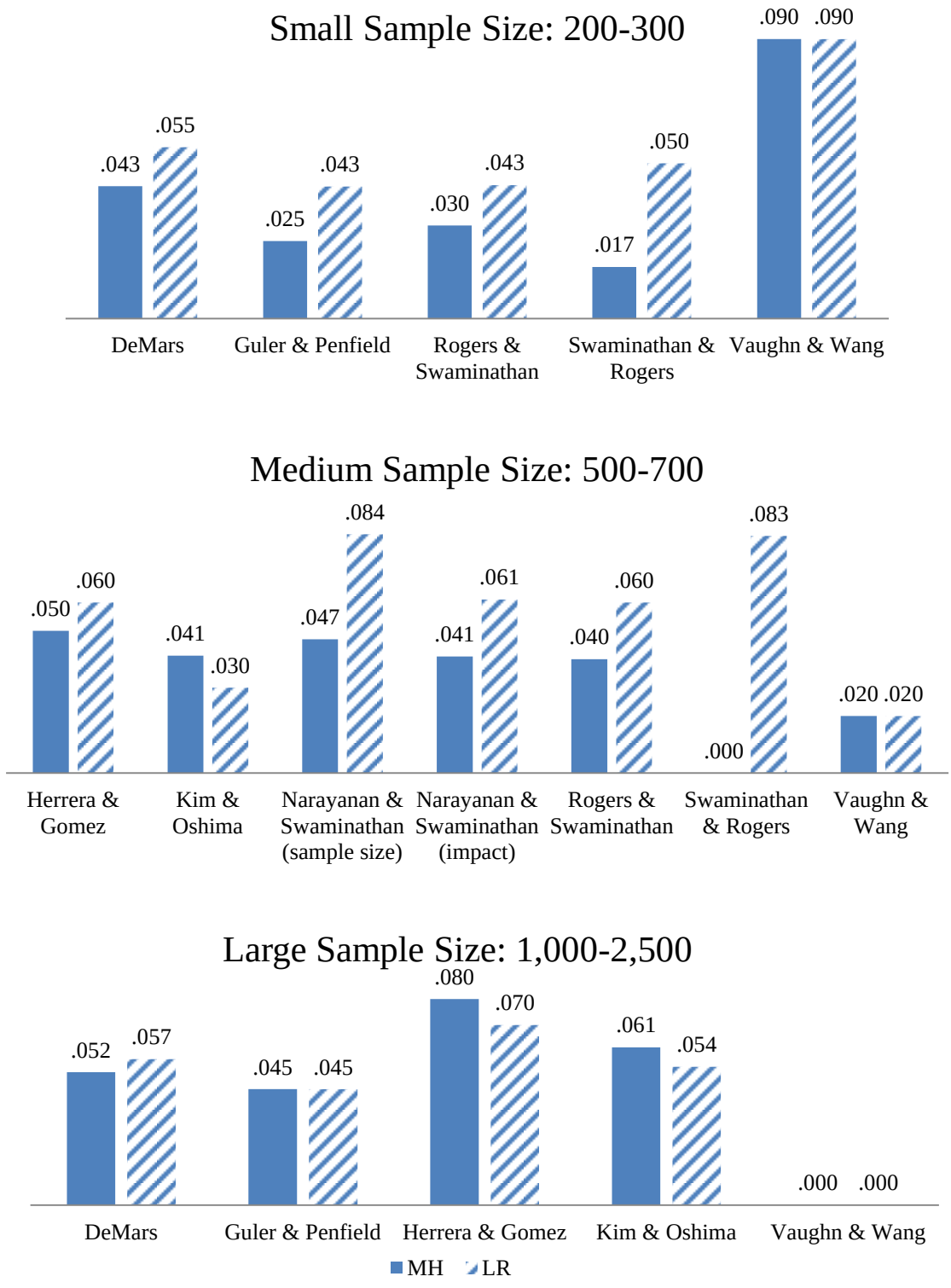


Figure 8. Type I Error Rate of Studies with Equal Sample Size and Impact = 0

Numbers atop the bars are the Type I error rates. All studies had nominal Type I error rate of .05.

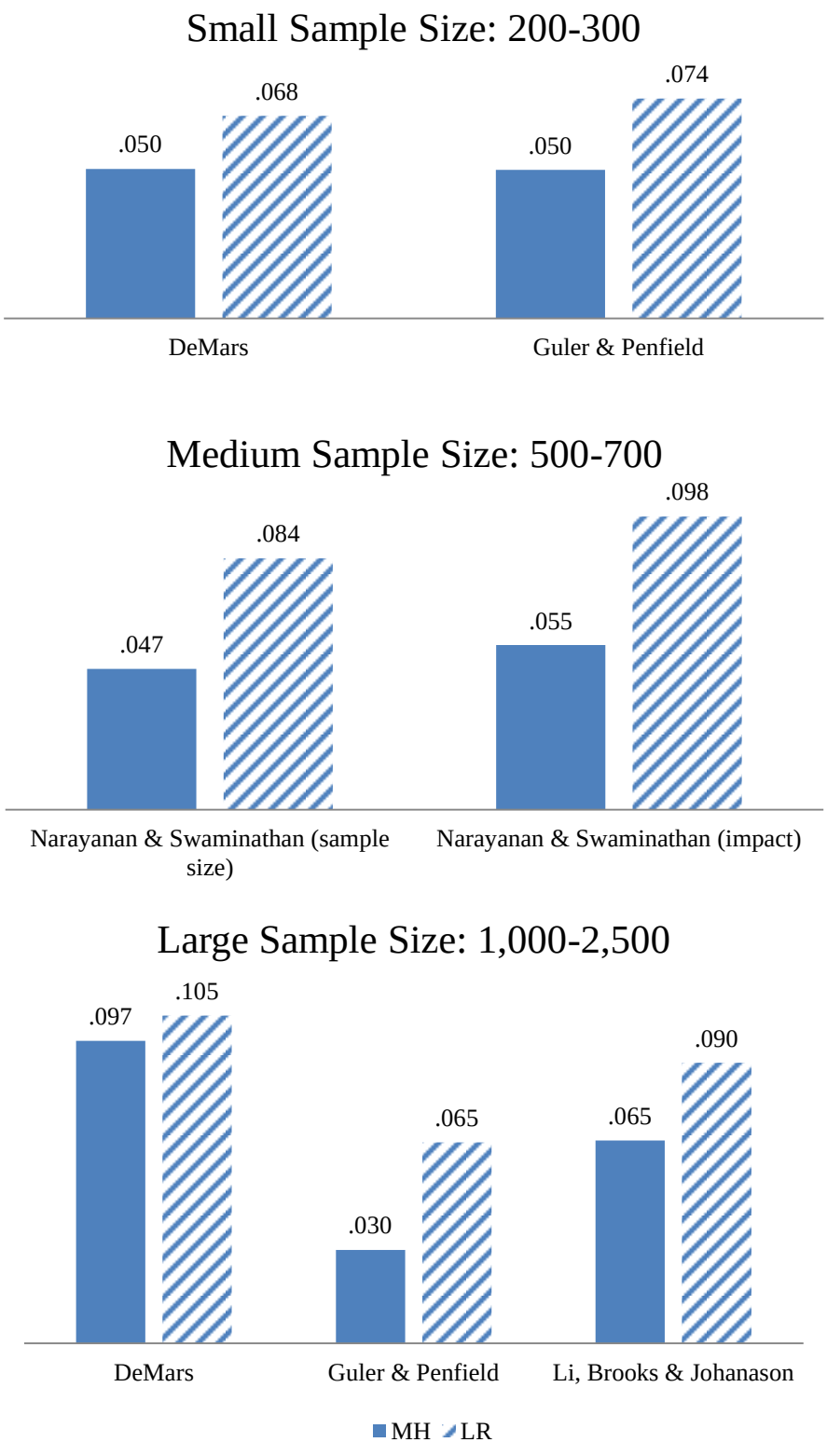


Figure 9. Type I Error of Studies with Equal Sample Size and Impact = 1

equal sample size conditions are presented in Figure 8. Figure 9 depicts congruent data for the impact equal to 1 condition.

Unequal sample size data fell into three categories: small/medium (200-300/500-700), small/large (200-300/1,000-2,500), and medium/large (500-700/1,000-2,500). Eight of the included studies contained unequal sample size data: five with small/medium (Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010), four with small/large (Güler & Penfield, 2009; Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010), and four with medium/large (de Ayala et al., 2002; Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010). While most unequal sample size data were usable Herrera & Gomez (2008) presented three groupings that fell below the small range for this study (100/500, 125/500, 167/500), one grouping that exceeded the range for medium (750/1,500) and one grouping that fell between the small and medium groupings (375/1,500). A detailed list of sample sizes for each study can be found in Appendix V. MH and LR Type I error data for unequal sample size data for impact equals zero and impact equals one are found in Figures 10 and 11, respectively.

Test Length

Test length was a constant for seven studies (de Ayala et al., 2002; Güler & Penfield, 2009; Herrera & Gomez, 2008; Li, Brooks & Johanson, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Vaughn & Wang, 2010). Three studies manipulated test length: one study used test lengths of 20 and 40 (Kim & Oshima, 2012), DeMars (2009) used three variations of test length (20, 40 and 60) and Swaminathan and Rogers (1990) overlapped with DeMars on the lengths of 40 and 60 while adding an 80 item test. Number of replications

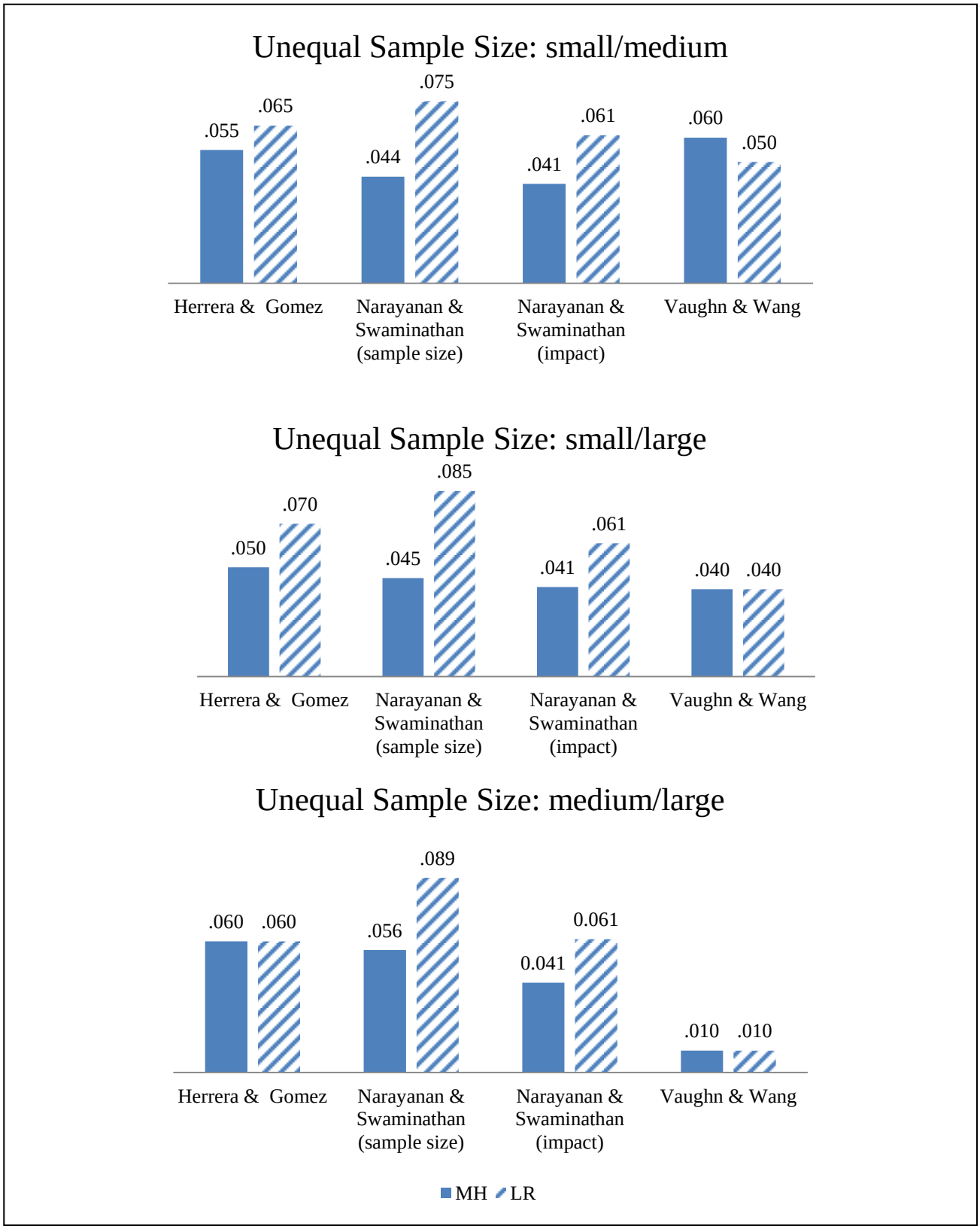


Figure 10. Type I Error Rates of Studies with Unequal Sample Size and Impact = 0

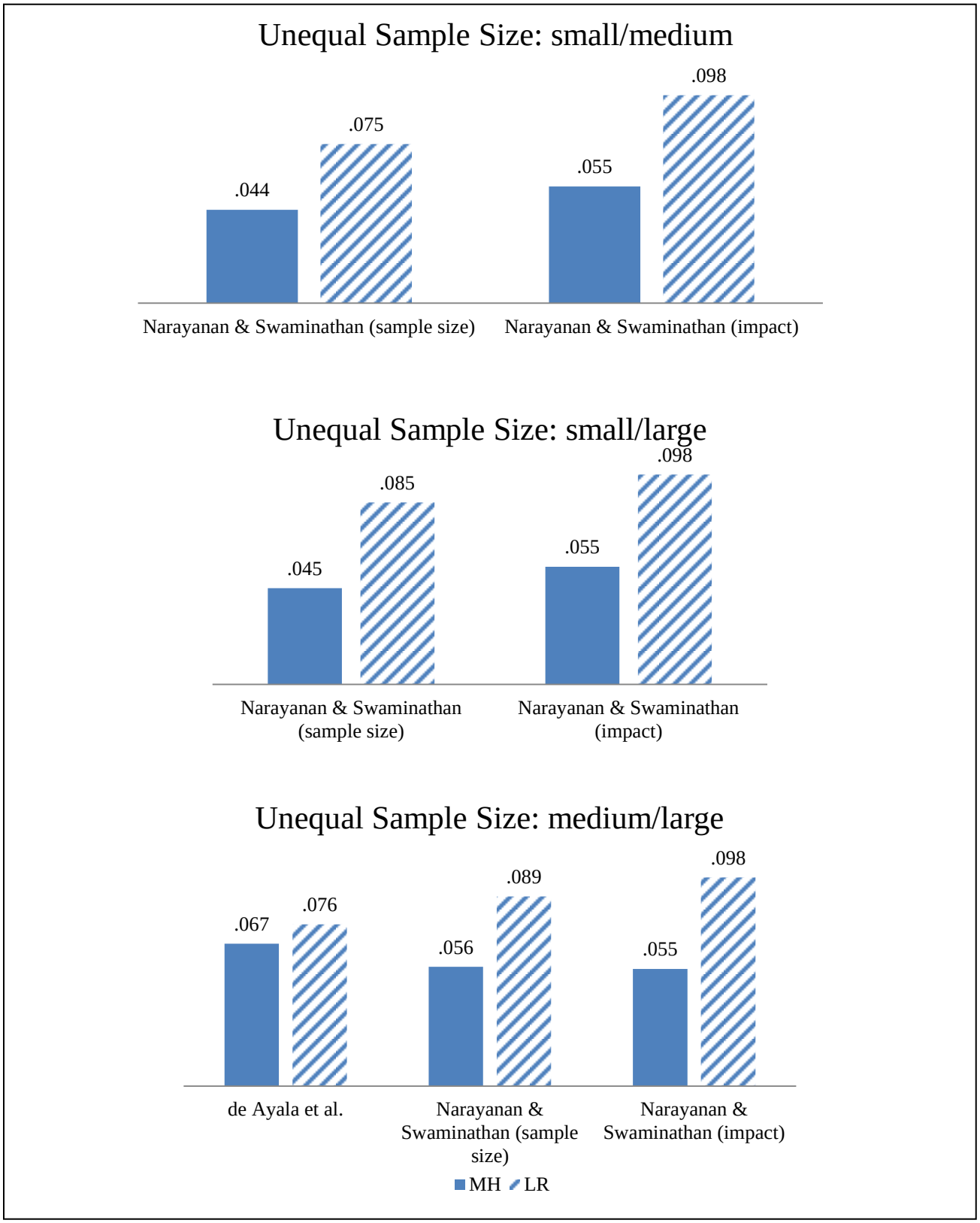


Figure 11. Type I Error Rates of Studies with Unequal Sample Size and Impact = 1

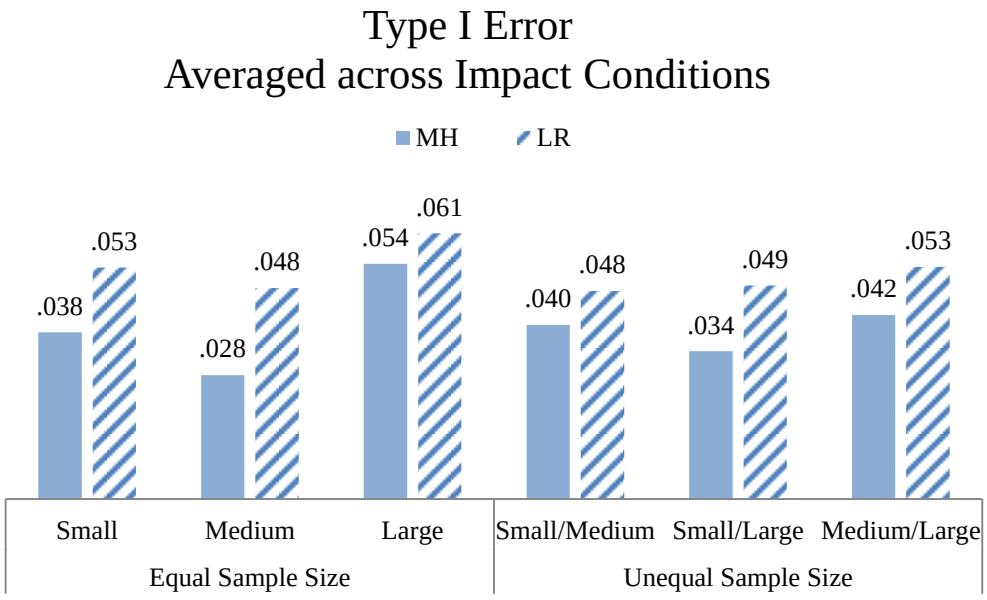
for the simulation studies varied from 20 (Swaminathan & Rogers, 1990) to 10,000 (Li, Brooks & Johanson, 2012) with 100 being the most commonly selected number (DeMars, 2009; Herrera & Gomez, 1008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). Number of replications and test length for each included study are shown in Appendixes E and W, respectively.

DIF Percentage

Another substantive study characteristic with the potential to affect Type I error is the percentage of DIF simulated for each test. DIF percentage varied from 0% to 30% among included studies. Two studies did not simulate DIF (Li, Brooks & Johanson, 2012; Rogers & Swaminathan, 1993), five studies exhibited static DIF percentages between 10% and 20% (Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) while three studies manipulated DIF percentage (de Ayala et al., 2002; DeMars, 2009; Narayanan & Swaminathan, 1996). Since disaggregated DIF percentage data were not available for DeMars (2009) that condition could only be compared for de Ayala et al. (2002) and Narayanan and Swaminathan (1996). For de Ayala et al. (2002) unequal impact of one, was the only ability distribution simulated; therefore, DIF percentage could only be compared for two studies and then only for the condition of impact equal to one. The percentage of DIF for each study is shown in Appendix U. Comparison of DIF percentage to other substantive study characteristics can be achieved through examination of the Final Coding Table in Appendix J.

Research question one compared Type I error performance for MH and LR using Type I error rates. Across all conditions and sample sizes the overall conclusion was that MH had the lowest Type I error rates.

Numbers atop the bars are the Type I error rates. All studies had nominal type I error of .05.



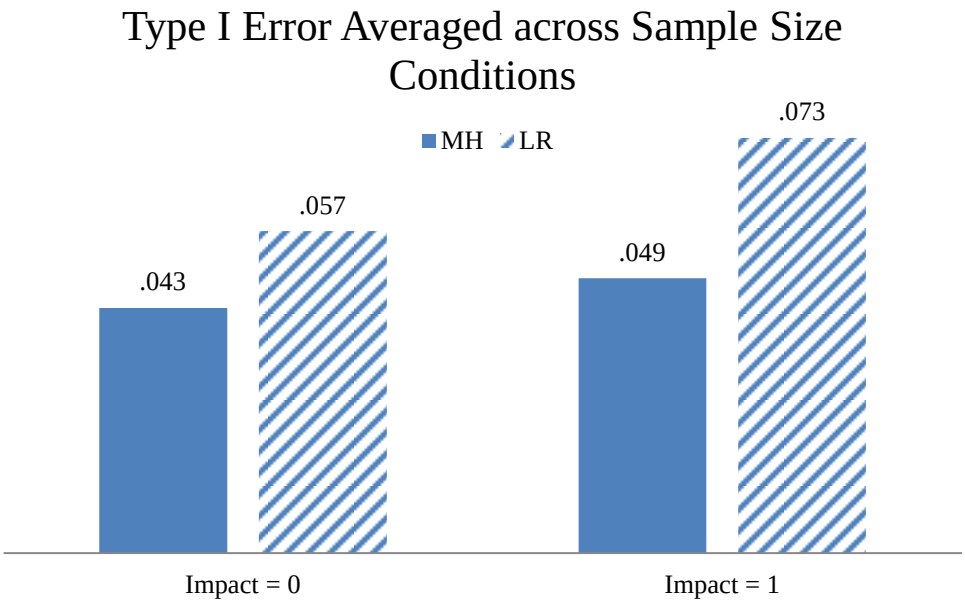


Figure 12. Type I Error Rates Averaged across Sample Size by Impact

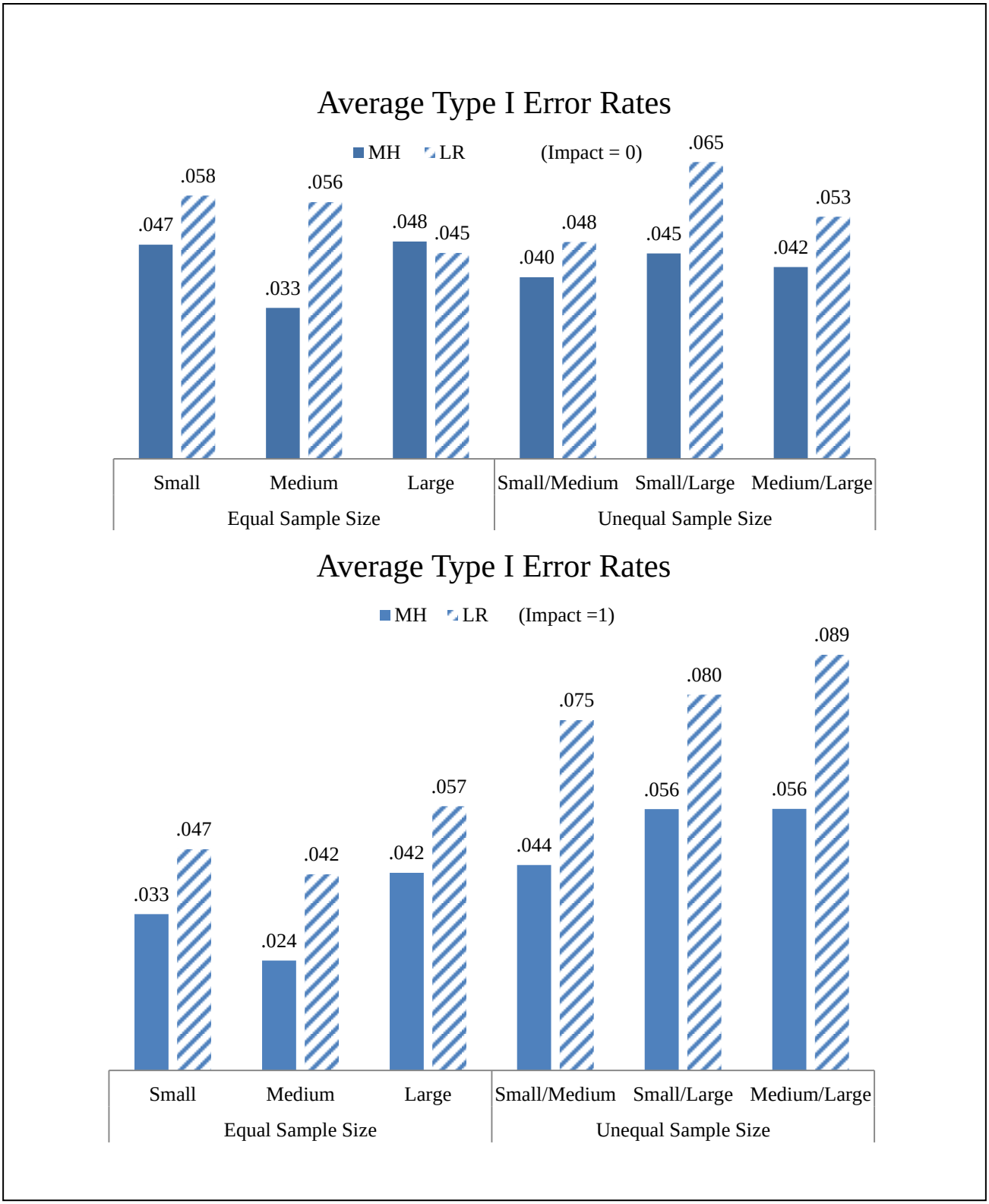


Figure 13. Type I Error Rates by Impact

Research Question 2:

Deviations from .05 Nominal Type I Error Rate by Condition for MH and LR

The data for Research Question 2 are organized in the same manner as for Research Question 1, so in the sections that follow graphics displaying the deviation of Type I error rates from the nominal .05 level for each compared condition in the included studies will be discussed. Deviation values were calculated by subtracting .05 from MH and LR Type I error values. This means that the deviation values are negative if the MH or LR values are less than .05 and positive if the MH or LR Type I error values exceed .05. Therefore higher bars in the graphics indicate greater Type I error values. Also, values below the x-axis are less than the nominal .05 Type I error rate while values above the x-axis represent conditions with Type I error rates greater than the .05 nominal value. Naturally it follows that values on the x-axis are those at the .05 level.

Impact and Sample Size

While the majority of studies presented data for the impact equals zero condition, three (DeMars, 2009; Güler & Penfield, 2009; Narayanan & Swaminathan, 1996) simulated equal and unequal ability distributions, and de Ayala et al. (2002) exhibited data only for impact equals one. Therefore, sample size was categorized as having equal (impact = 0) or unequal ability differences (impact = 1). As in research question one, sample size was divided into equal and unequal categories determined by focal and reference group size with each of those clusters being further subdivided into three size-related groups. Thus, organization of data displaying deviations from the nominal Type I error for MH and LR is presented across six categories for equal sample size (impact=0 & impact=1 x small, medium or large), shown in Figures 13 and 15,

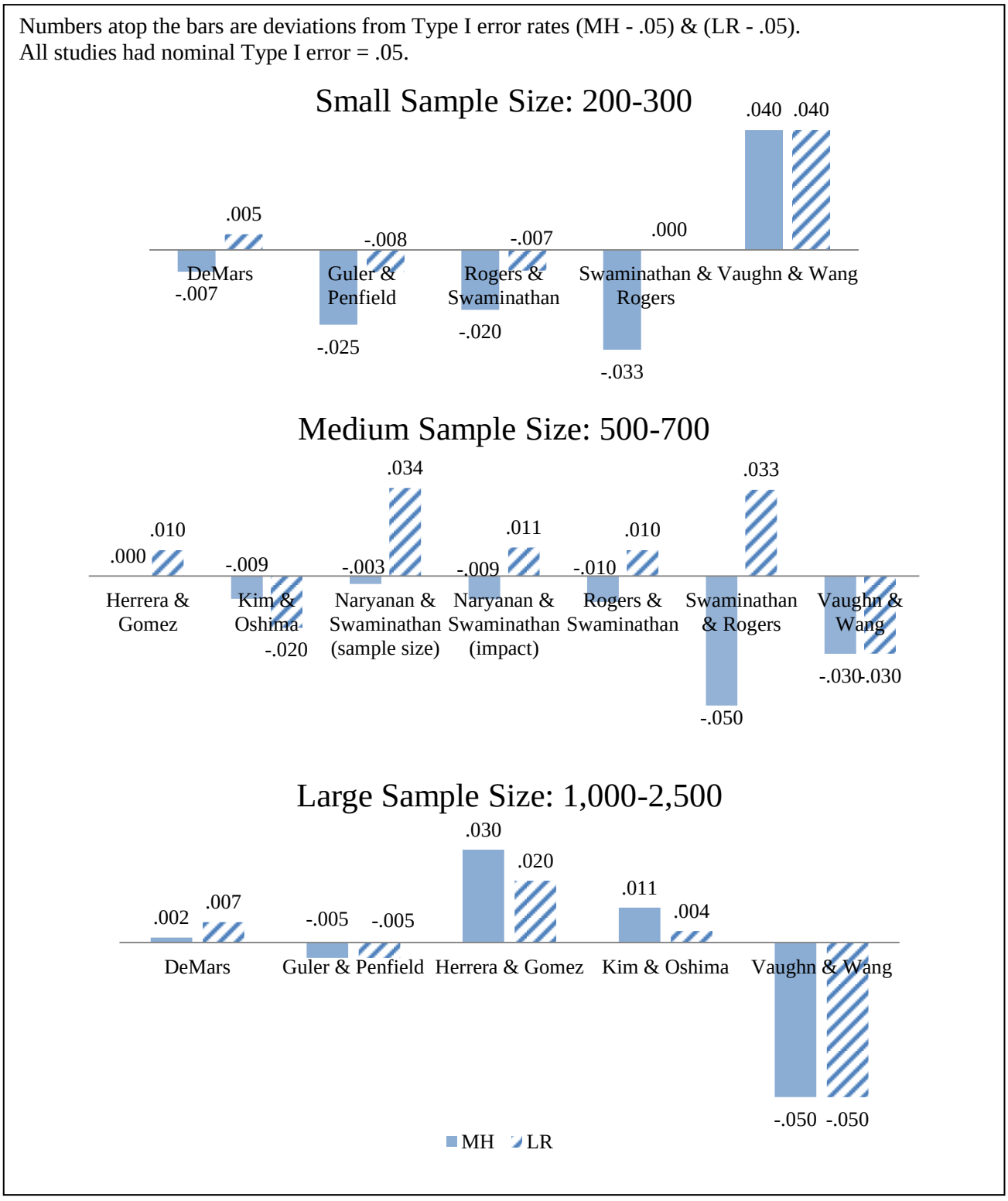


Figure 14. Type I Error Rate deviation from .05 for Studies with Equal Sample Size and Impact = 0

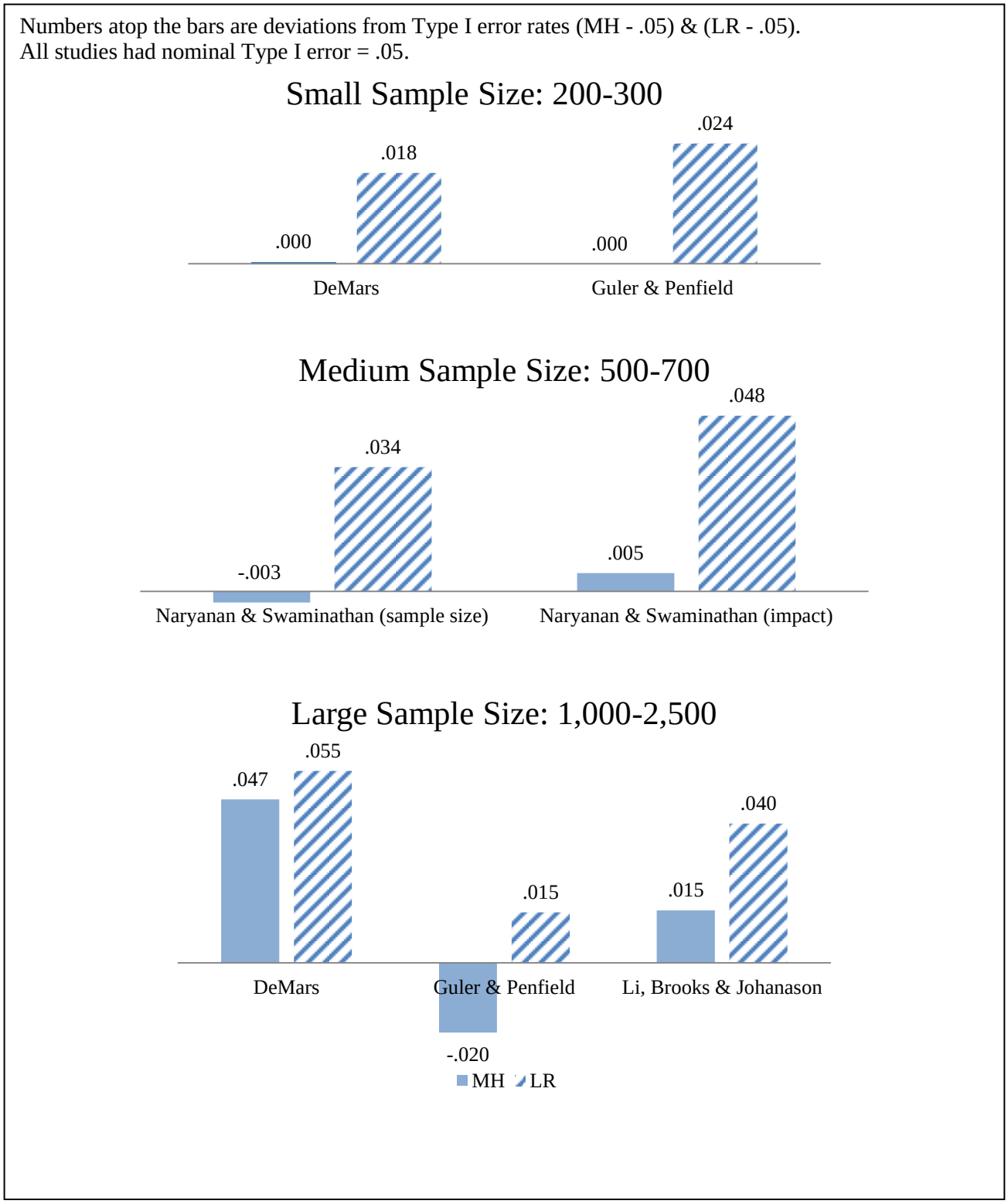


Figure 15. Type I Error Rate deviation from .05 for Studies with Equal Sample Size and Impact = 1

Numbers atop the bars are deviations from Type I error rates (MH - .05) & (LR - .05).
All studies had nominal Type I error = .05.

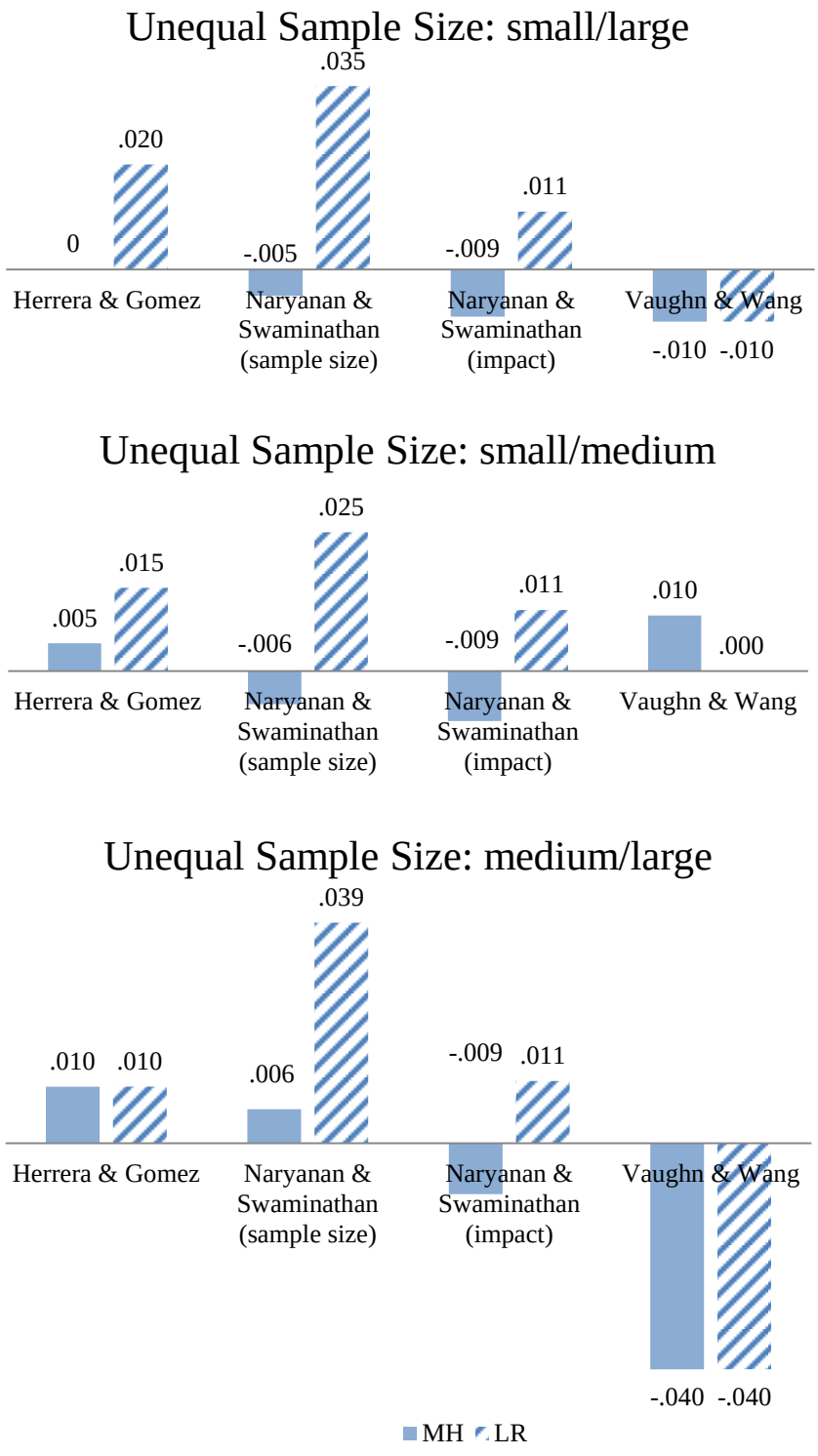


Figure 16. Type I Error Rate deviation from .05 for Studies with Unequal Sample Size & Impact = 0

Numbers atop the bars are deviations from Type I error rates (MH - .05) & (LR - .05).
All studies had nominal Type I error = .05.

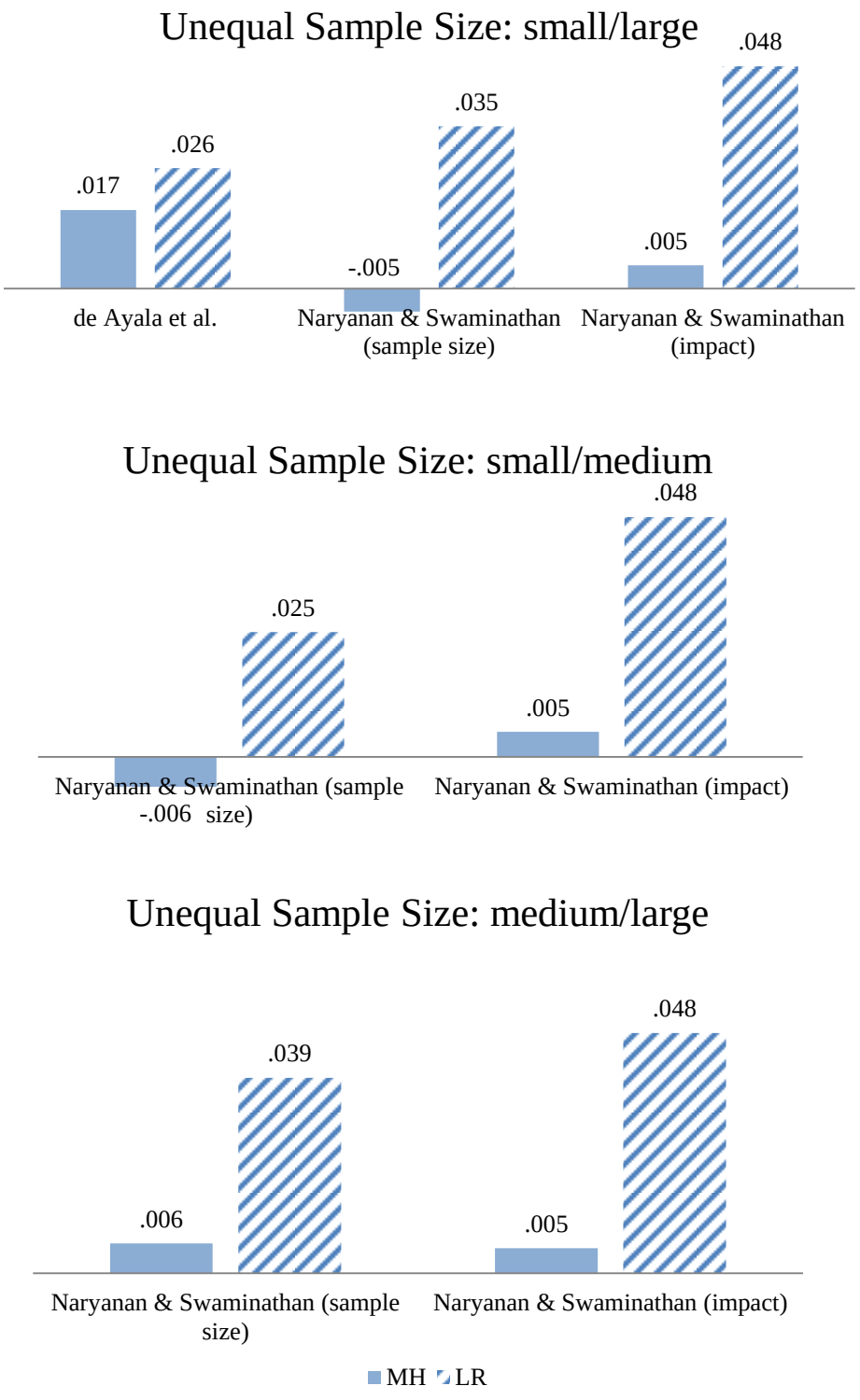


Figure 17. Type I Error Rate deviation from .05 for Studies with Unequal Sample Size & Impact = 1

and six categories for unequal sample size (impact=0 and impact=1 x small/medium, small/large, medium/large), shown in Figures 16 and 17. Numerical ranges for sample size are shown in Appendix X.

Test Length and Replications

Test Length. Grouping of test length for analysis of deviation from nominal Type I error rate was conducted in the same fashion as for research question 1. Test length was divided into three groups: short (20-30), moderate (40-60), and long (80-100). Three studies simulated tests that were considered short in length (de Ayala et al., 2002; DeMars, 2009; Kim & Oshima, 2012). The condition of a short test with an impact of one, encompassed two studies (de Ayala et al., 2002; DeMars, 2009), and comparing short tests with an impact of zero also pertained to two studies (DeMars, 2009; Kim & Oshima, 2012). All included studies except two (de Ayala et al., 2002; Herrera & Gomez, 2008) simulated tests of moderate length, and seven (DeMars, 2009; Guler & Penfield, 2009; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) of those were compared for the impact equal to zero condition. The condition of impact equal to one and moderate test length was compared for three studies (DeMars, 2009; Guler & Penfield, 2009; Li, Brooks & Johanson, 2012). Only two (Herrera & Gomez, 2008; Swaminathan & Rogers, 1990) studies simulated long tests both which had an impact of zero.

Replications. Replications were organized into three categories: small (20-50), medium (100-300), and large (1,000–10,000). Swaminathan & Rogers (1990) & de Ayala et al. (2002) used a small number of replications, though these studies found common ground with a DIF percentage of 20%, they differed with regard to other substantive study characteristics. Six studies (DeMars, 2009; Guler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012;

Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) conducted simulations using the medium number of replications. Two studies (Li, Brooks & Johanson, 2012; Vaughn & Wang, 2010) used a large number of replications, though these studies shared a moderate test length (40-60) and large, equal sample size (1,000/1,000), they differed on the conditions of impact (1/0) and DIF percentage (0%/20%), respectively.

DIF Percentage

Ranges for DIF percentage were none (0%), low (10-15%), moderate (20%), and high (30%). These can be found in tabular form in Appendix X. Four studies simulated the condition of no DIF: two (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) with impact equal to zero and two (de Ayala et al., 2002; Li, Brooks & Johanson, 2012) with impact equal to one. The low DIF percentage label (10-15%) applied to six studies, and was compared for impact equal to zero (DeMars, 2009; Guler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996) and impact equal to one (de Ayala, et al., 2002, DeMars, 2009, Guler & Penfield, 2009; Narayanan & Swaminathan, 1996). Examination of Appendix U reveals that DeMars (2009) used three levels of DIF percentage (10%, 15%, 30%), however, since disaggregated data were not available, DIF percentage results from her study were averaged with low DIF percentage results because the majority of the data fit that label. The case of moderate DIF (20%) warranted comparison of three studies (Narayanan & Swaminathan, 1996; Swaminathan & Rogers; Vaughn & Wang, 2010) with equal ability distributions and two with unequal ability distributions (de Ayala et al, 2002; Narayanan & Swaminathan, 1996). Since Narayanan & Swaminathan provided Type I error results for all substantive study characteristics though not in disaggregated form, results from their study

appear as two different values for each compared substantive study characteristic. A detailed list of all substantive study characteristics can be found in the Final Coding Table of Appendix J.

Research Question 2 examined deviations from the nominal .05 Type I error rate. In the preceding graphics bars on the x-axis with a value of zero have a Type I error rate equal to the nominal .05 rate, those extending above the x-axis exceed the nominal .05 level and those with bars extending below the x-axis have a value below the nominal .05 level. Since Type I and Type II errors are linked, values of Type I error equal to .05 or slightly below .05 are desirable because those values demonstrate control of Type I error without impeding power, or correct identification of DIF items. For this research question MH is also the recommended statistical method for control of Type I error because it generally displays lower Type I error rates than LR.

Research Question 3

Analyzing the constituent studies according to substantive study characteristics required preliminary exploratory work. Initially, bar graphs of all substantive study characteristics were created individually. However, during that process impact and sample size were shown to be particularly predictive of Type I error effect size. Therefore, analysis of additional substantive study characteristics was conducted on data that were organized either by impact or sample size or both in order to clarify the comparison.

As mentioned above the effect size measure used to compare included studies was Type I error effect size. Type I error effect size values for *d*' are shown in Appendix K alongside the steps carried out to calculate the proportion. Calculating the Type I error effect size for LR and MH set the back drop for the creation of groups of graphics shown in Figures 18 through 23

which summarize the relationships between substantive study characteristics.

Comparison of articles was broken down by ranges of substantive study characteristics which are shown in Appendix X. Ranges were established by comparing the characteristic of interest across studies using the Final Coding Table from Appendix J. The following paragraphs discuss the categorization of studies according to the four substantive study characteristics: impact, sample size, percentage of DIF, and test length.

Impact and Sample Size

Though existence of DIF is only possible within equal ability groups, the presence of examinees of varying ability levels within the pool of testing candidates can complicate the correct identification of unbiased items. DIF is used to refer to a situation where test items perform differently for examinees of equal ability, while impact describes the situation where test items discriminate between examinees of differing ability levels. Cases of no ability difference used means of zero and standard deviations of one for both focal and reference groups (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Li, Brooks & Johanson, 2012; Narayanan & Swaminathan, 1996, Rogers & Swaminathan, 1990; and Vaughn & Wang, 2010). Most studies simulating unequal impact used means of zero for the reference group and means of negative one for the focal group with a standard deviation of one for both groups (de Ayala et al., 2002; DeMars, 2009; Güler & Penfield, 2009; Narayanan & Swaminathan, 1996). Two studies manipulated impact utilizing values too unique for comparison (DeMars 2009; Vaughn & Wang, 2010). Appendix T provides detailed impact data for each study. Since disaggregated Type I error data were not available for Narayanan and Swaminathan (1996) Type I error effect size measures are shown separately for impact and sample size for that study in figures 18 through 23.

Sample size refers to the number of candidates or examinees present in each of the testing groups, reference and focal. Sample size can be equal, meaning that both groups contain the same number of examinees, or unequal, where the focal group is usually smaller and the reference group is larger. Though some DIF identification methods tend to perform better when focal and reference groups are equal in number, the focal group is typically smaller, making unequal reference and focal groups a more realistic scenario. Ranges of sample size for each study are shown in Appendix X.

Equal sample size and impact. Equal sample size was analyzed in conjunction with impact resulting in two categories: impact of zero and impact of one. Studies were further subdivided into small (200-300), medium (500-700), and large (1,000-2,500) groups. Therefore six categories of impact and equal sample size were created: 1) impact of zero with equal small sample size (DeMars, 2009; Güler & Penfield, 2009; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan and Rogers, 1990; Vaughn & Wang, 2010), 2) impact of zero with equal medium sample size (Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan and Rogers, 1990; Vaughn & Wang, 2010), 3) impact of zero with equal large sample size (DeMars, 2009; Güler & Penfield, 2009; Kim & Oshima, 2012; Vaughn & Wang, 2010), 4) impact of one with equal small sample size (DeMars, 2009; Güler & Penfield, 2009; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010), 5) impact of one with equal medium sample size (Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010), and 6) impact of one with equal large sample size (DeMars, 2009; Güler & Penfield, 2009; Li, Brooks & Johanson, 2012; Vaughn & Wang, 2010). Comparison of Type I error effect size for studies with equal sample size and impact of zero are shown in Figure 17 while those with impact of one are shown in Figure 18.

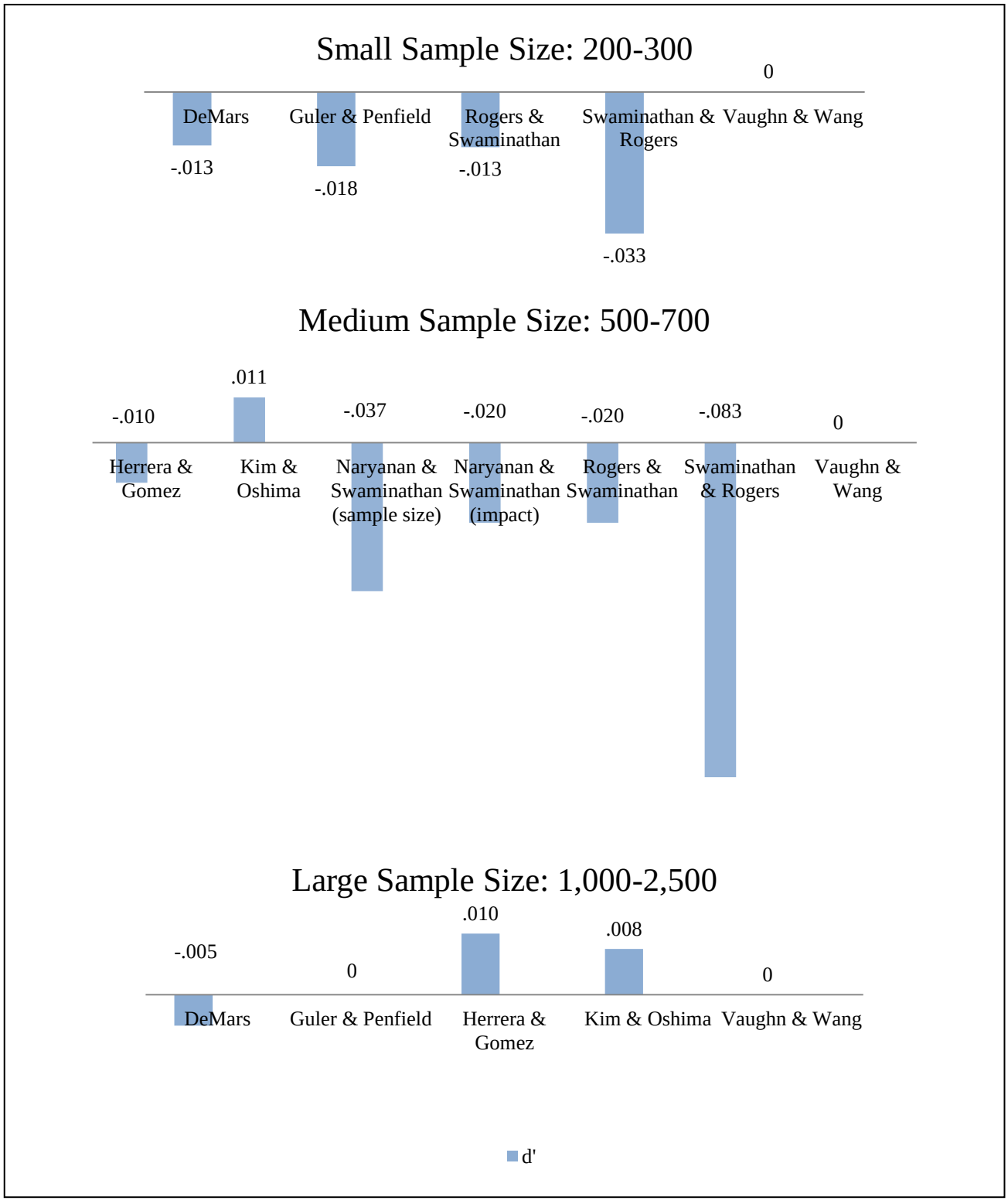


Figure 18. Type I Error Effect Size of Studies with Equal Sample Size and Impact = 0

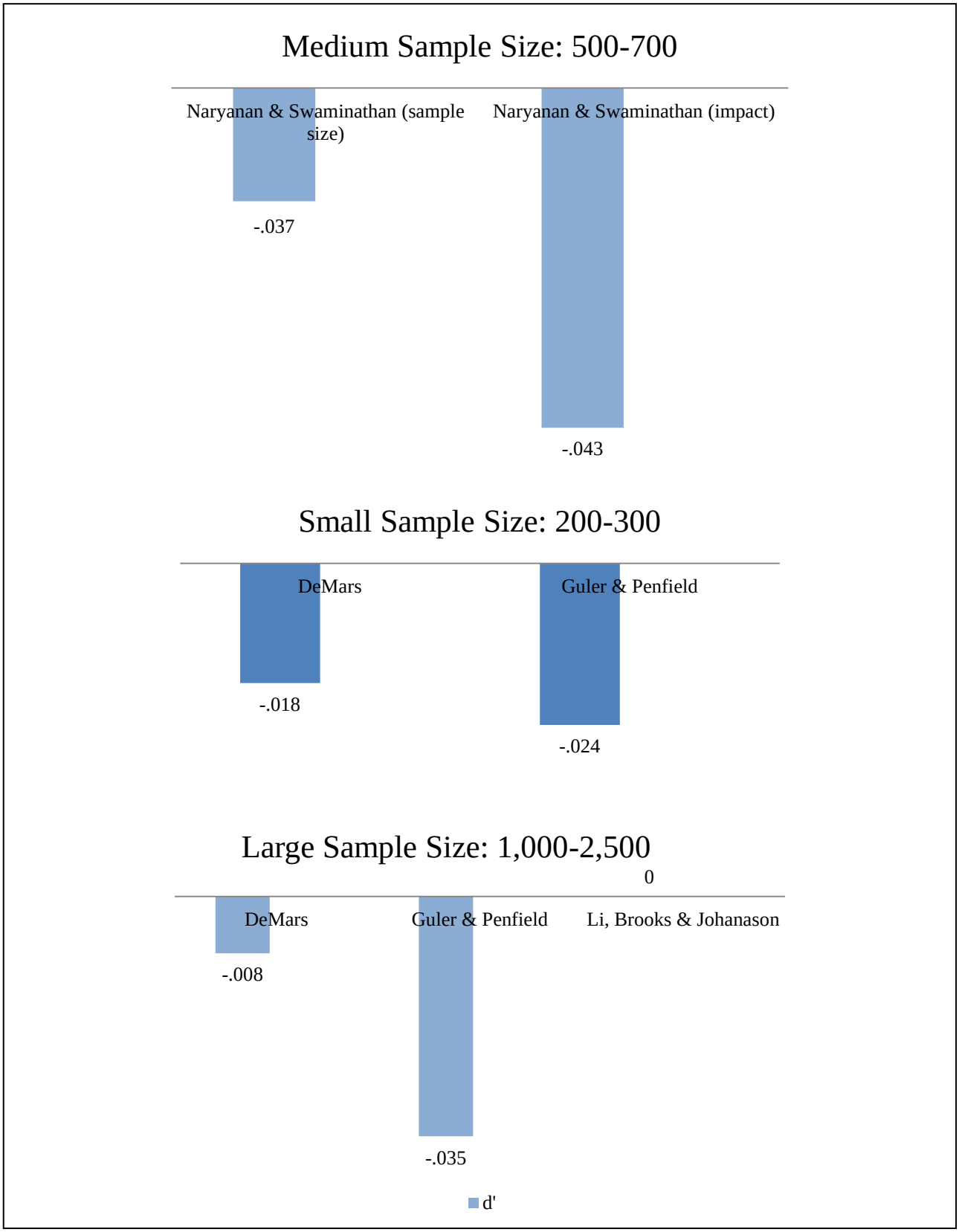


Figure 19. Type I Error Effect Size of Studies with Equal Sample Size and Impact = 1

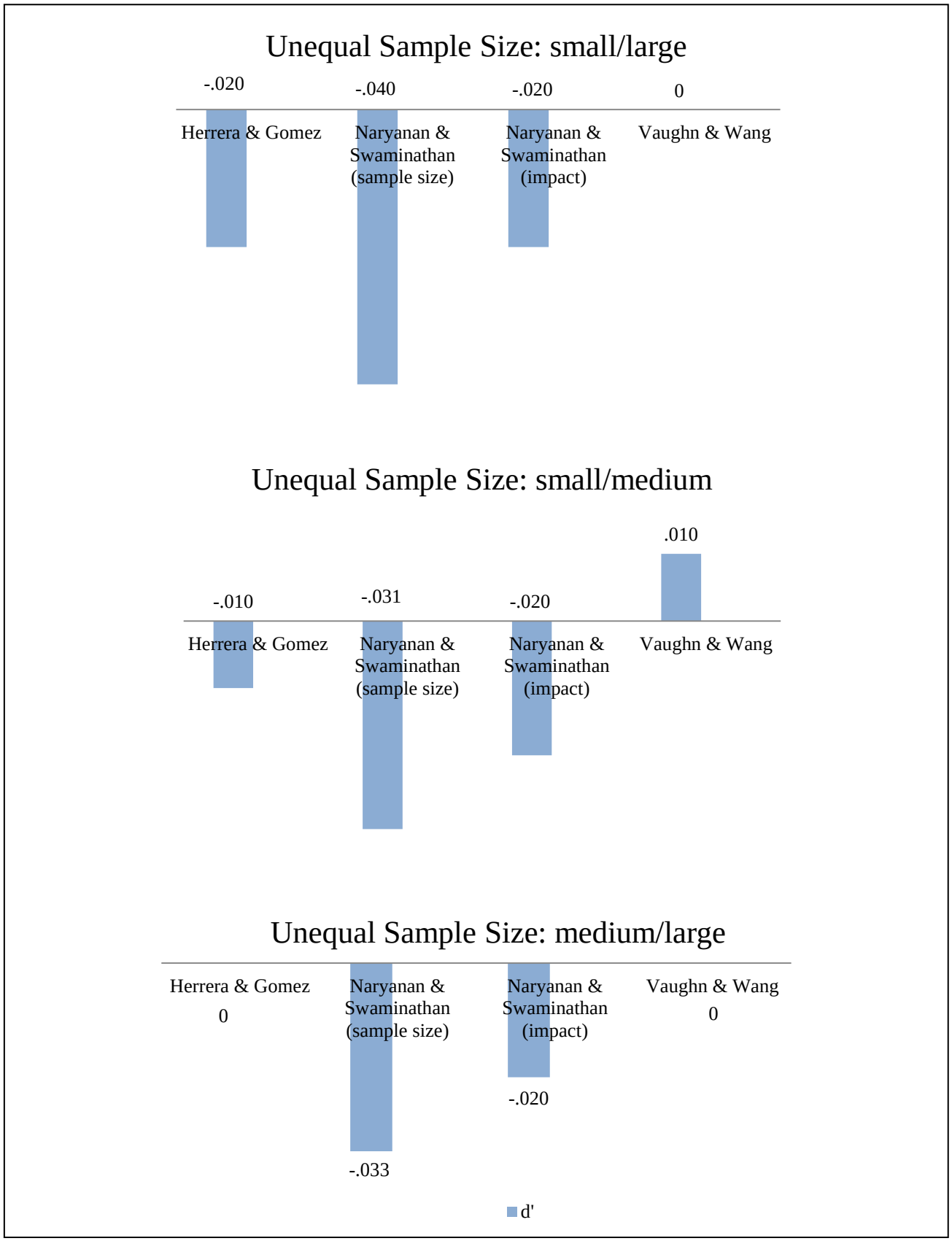


Figure 20. Type I Error Effect Size of Studies with Unequal Sample Size & Impact = 0

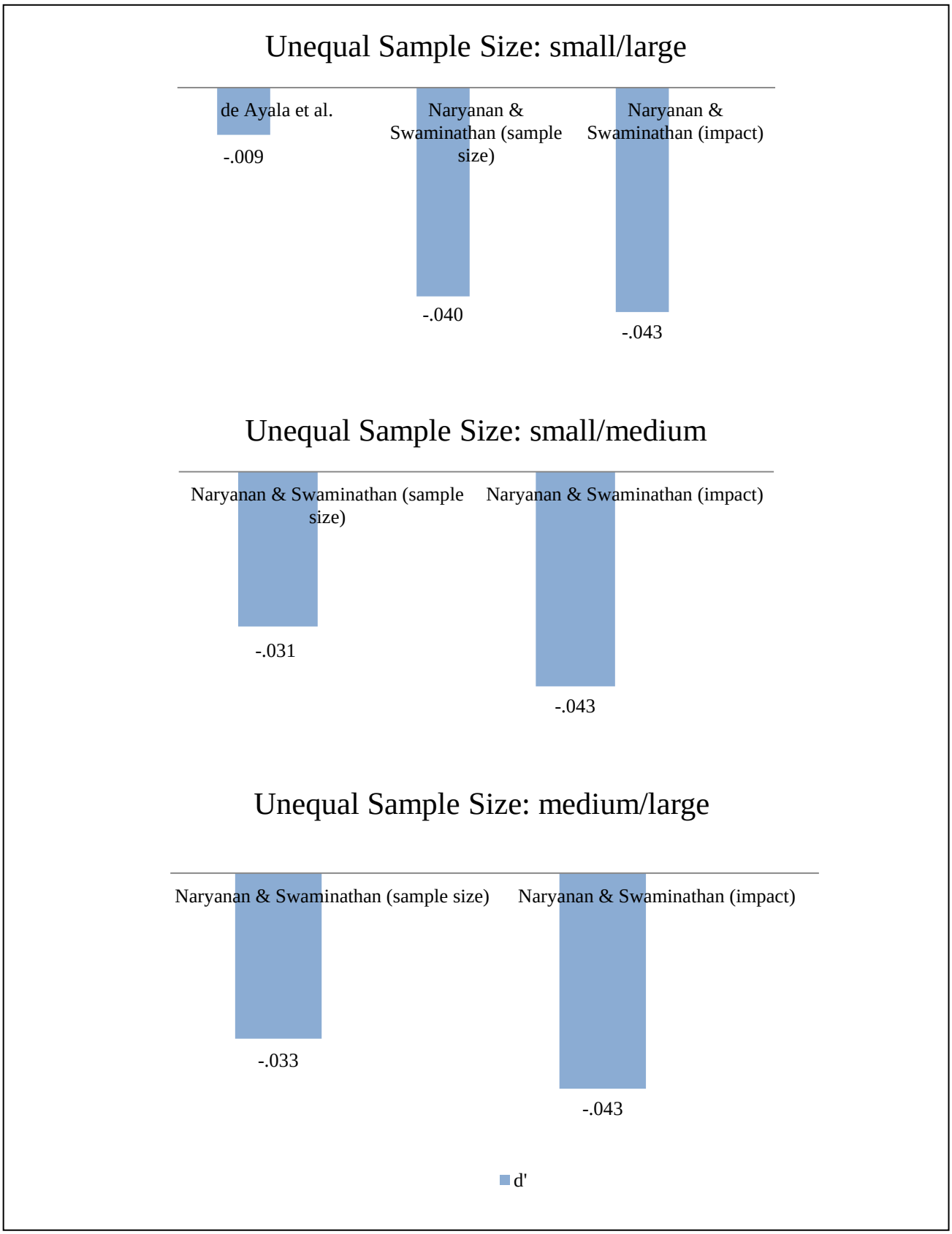


Figure 21. Type I Error Effect Size of Studies with Unequal Sample Size & Impact = 1

Unequal sample size and impact. Unequal sample size was also analyzed with impact resulting in two divisions of unequal sample size: impact of zero, shown in Figure 19 and impact of one, shown in Figure 20. Studies were further subdivided into small/medium (200-300/500-700), small/large (200-300/1,000-2,500), and medium/large (500-700/1,000-2,500) groups. Therefore six categories of impact and equal sample size were created: 1) impact of zero and unequal small/medium sample size (Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan and Rogers, 1990; Vaughn & Wang, 2010), 2) impact of zero and unequal small/large sample size (Güler & Penfield, 2009; Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010), 3) impact of zero and unequal medium/large sample size (Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010), 4) impact of one and unequal small/medium (Narayanan & Swaminathan, 1996), 5) impact of one and unequal small/large (Güler & Penfield, 2009; Narayanan & Swaminathan, 1996), and 6) impact of one and unequal medium/large (de Ayala et al., 2002; Narayanan & Swaminathan, 1996). Some ranges from 500/100-250 and 1500/300-750 (Herrera & Gomez, 2008) were too unique for comparison. A table detailing the sample size ranges for each study can be found in Appendix X. A graphical comparison of studies with unequal sample size and impact of zero is shown in Figure 19. While studies having unequal sample size and impact of one are compared in Figure 20.

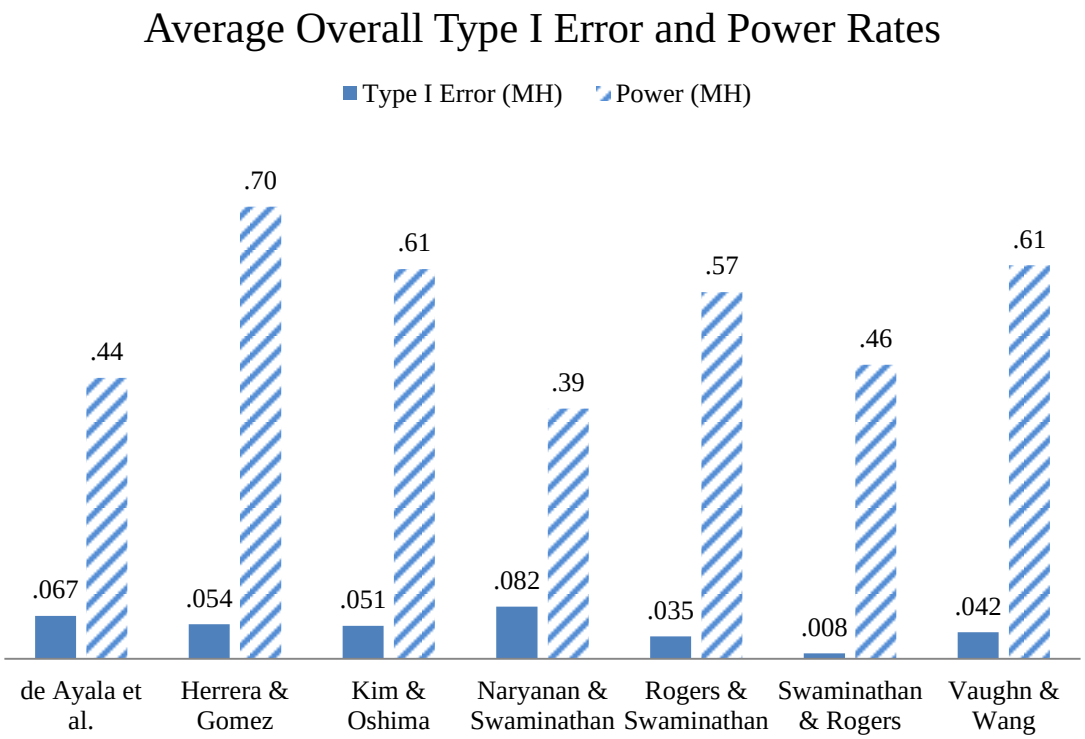
Figure 22. Type I Error Effect Size Averaged across Impact & Sample Size

Sample Size and Impact Averaged Across Studies. To compare sample size and impact the Type I error effect size for these substantive study characteristics was averaged across studies. The resulting graphics are found in Figures 21 and 22. Studies of equal sample size falling into categories of small (DeMars, 2009; Güler & Penfield, 2009; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1993; Vaughn & Wang, 2010), medium (Herrera & Gomez, 2008 Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) and large (DeMars, 2009; Güler & Penfield, 2009; Herrera & Gomez, 2008; Kim & Oshima, 2012; Vaughn & Wang, 2010) with impact equal to zero are shown in Figure 18. Though four studies (DeMars, 2009; Güler & Penfield, 2009; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010) had small, equal sample size and impact of one, only one study fell into the medium (Narayanan & Swaminathan, 1996) and large (Li, Brooks & Johanson, 2012) category. Sample size conditions averaged across impact are shown in Figure 22.

In addition to the comparison of deviation scores of MH and LR from the .05 nominal Type I error rate, a meta-analysis software package, MetaAnalyst, was used to compare deviation scores, test for the significance of the random effects model and produce a forest plot. Results, shown in appendixes Y and Z, indicated that MH was the preferred method for seven out of ten studies. Though results were not significant, a shift of only one thousandth of a point would have turned the tables. Also, results testing for significance with respect to the random effects model did show significance confirming that the random effects model was appropriate for this data set.

Research Question 3 displayed the Type I error effect size as *d'*. Since it was calculated by subtracting LR from MH (MH – LR), negative values indicated that MH performed better

Numbers atop the bars are the Type I error rates & power percentages. All studies had nominal Type I error rate of .05.



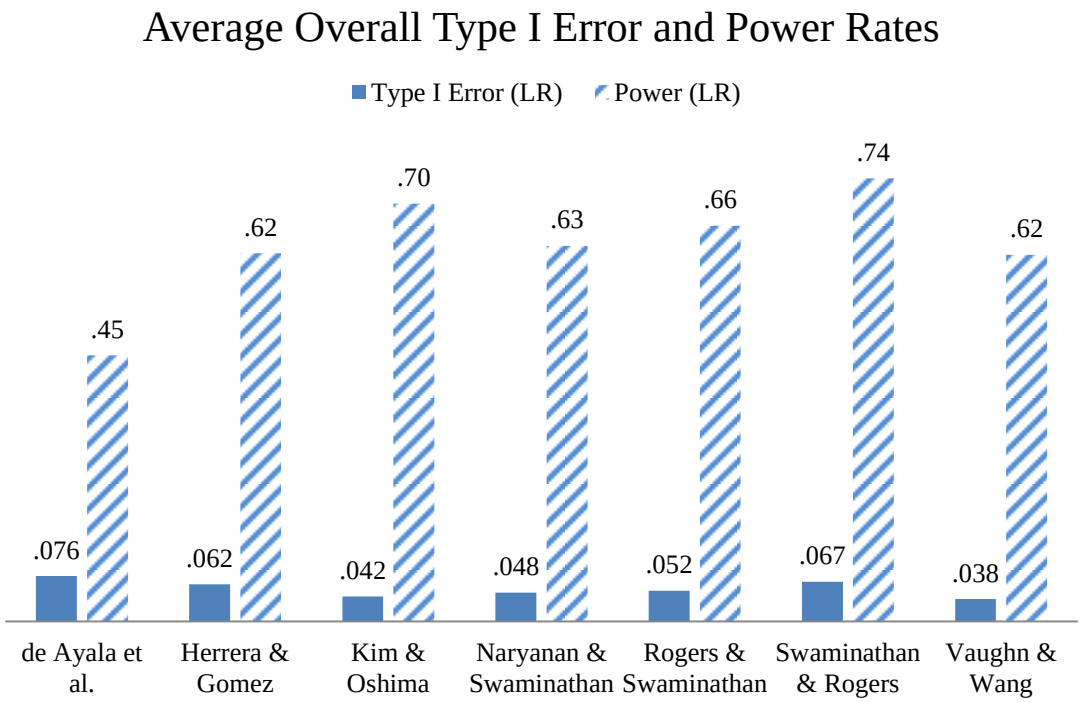
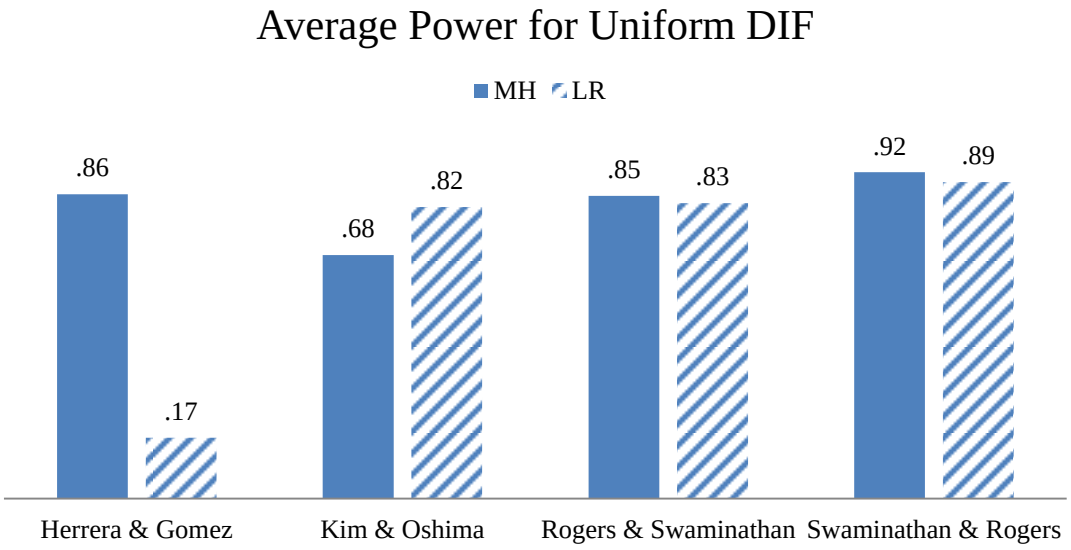


Figure 23. Comparison of Type I (false positive) error rates & Power (correct identification) for MH & LR by study

Numbers atop the bars are the power rates.



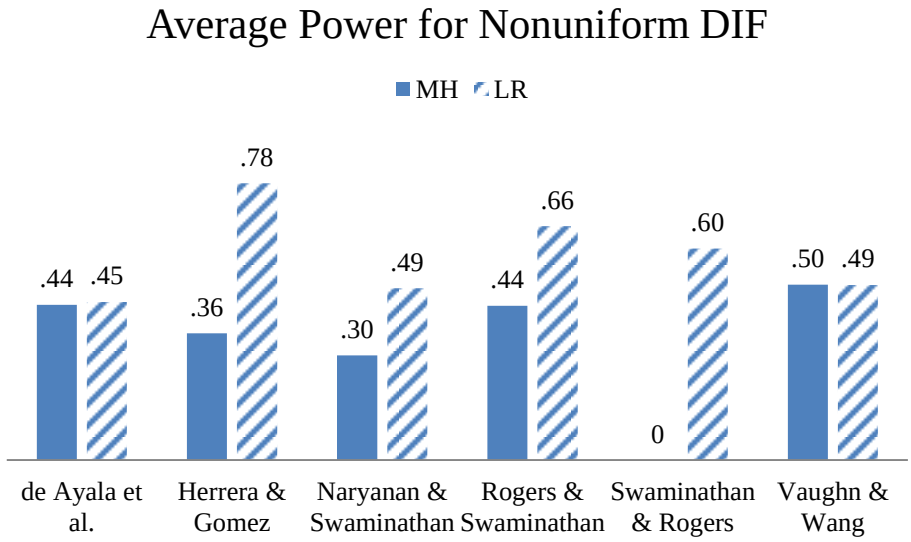


Figure 24. Comparison of Power (correct identification) for MH & LR by study for uniform and nonuniform DIF.

with respect to Type I error control. Therefore, in regard to research question 3, practitioners desiring to control Type I error would likely prefer MH for identification of DIF items.

Research Question 4

Though Type II error is as important as Type I error, the inclusion criteria for this meta-analysis only specified the presence of Type I error. However, since seven included studies, shown in Appendix G, provided Type II error power data for these studies were added post hoc. Type I error rates, also called false positives, occur when an event is recorded as happening even though it did not occur.

In the context of this study Type I error means than an item that does not contain DIF was identified as DIF-containing. In practice this mistake could cause unbiased test items to be removed from a test causing new items to be developed unnecessarily. On the other hand, Type II errors occur when an event happens yet it is not recorded as having occurred. Here, that would that an item containing DIF was not identified as DIF-containing. Such an error could result in a law suit if an examinee questioned the fairness of a test item, and the item was discovered to contain DIF. Therefore, in order for practitioners to their best at producing a test that is truly fair, Type I and Type II errors must be examined in tandem. The opportunity to conduct such an analysis is provided by Figure 23. Of the seven studies displaying power data, which are shown in Appendix G, five compared MH and LR with uniform DIF and six with nonuniform DIF.

Research Question 4 displayed power (1- Type II error) data adjacent to Type I error data. The trend became bifurcated with this question because with respect to power, LR is the preferred statistical method for identification of DIF items with nonuniform DIF, while MH is preferred for the situation of uniform DIF. Therefore, in regard to research question 4, for

practitioners prioritizing correct identification of DIF items over avoidance of false positive identifications, the nature of DIF, uniform or nonuniform, should be considered.

CHAPTER 5: DISCUSSION

This meta-analysis was driven by four research questions. Three were pre-determined while the fourth emerged during the process of carrying out the study. Before summarizing the results for each research question separately, an overview of meta-analysis procedures will be presented.

The original intent was to compare four substantive study characteristics: impact, sample size, DIF percentage and test length. However, the manner in which data were collected and presented in each of the 10 included studies necessitated changing the original plan. Percentage of DIF was divided into three ranges: none (0%), low (10-15%), moderate (20%), and high (30%), which are shown in Appendix X. Though percentage of DIF was reported for each study, only three studies (de Ayala et al., 2002; DeMars, 2009; Narayanan & Swaminathan, 1996) treated DIF percentage as a variable. Of those only de Ayala et al. (2002) disaggregated the data in a manner that allowed comparison. Therefore, DIF percentage was not compared across studies as a substantive study characteristic.

Test length was divided into three levels: short (20-30), moderate (40-60), and long (80-100), shown in Appendix X. Test length was treated as a variable for three studies (DeMars, 2009; Kim & Oshima, 2012; Swaminathan & Rogers, 1990). DeMars (2009) provided data disaggregated by test lengths of 20, 40 and 60 items. Kim and Oshima (2012) simulated test lengths of 20 and 40, which spanned the short and moderate test length ranges from Appendix X, and Swaminathan and Rogers (1990) used lengths of 40, 60 and 80 which spanned the moderate and large groups. Since only one category of test length, 40, was available for four studies, comparison of this substantive study characteristic was not completed as initially planned.

Because data for impact and sample size were available for each of the 10 included studies Type I error rates, Type I error deviation scores, and Type I error effect sizes were compared across studies for these substantive study characteristics. Before summarizing the results for each research question separately, an overview discussion integrating the results from the four research questions is presented.

Type I error rate data are presented in two different ways in research questions one and two. For comparisons of MH and LR statistical methods, graphics associated with research question one provide the best visuals. Since the bars displaying Type I error rates for MH are the smallest in most instances, research question one demonstrates the superior ability of MH over LR for controlling Type I error. The accepted rate of false positives; that is, Type I error, is .05. Research question two graphics provide the best juxtaposition of studies with Type I error rates at, above or below this nominal rate. In Figures 13 through 16, bars above the x-axis indicate studies with Type I error rates above the nominal .05 rate, bars below the x-axis show studies whose rates are below the nominal rate and studies with a zero value for Type I error deviation scores are those with Type I error rates equal to the .05 nominal rate.

Graphics associated with research questions three and four provide visuals for Type I error effect size and power, respectively. The Type I error effect size shown for each study in the graphics for research question three was calculated by subtracting MH proportion values from LR (MH – LR); a worked example for each study is shown in Appendix K. Therefore, if the effect size is negative LR Type I error rates were larger than MH, meaning that MH controlled Type I error best. If the effect size is zero then MH and LR Type I error rates were equal. Positive Type I error effect size values indicate that LR controlled Type I error best. The distance the bars extend from the x-axis indicates the difference between the Type I error rates for the two

statistical methods, MH and LR. Across all 56 Type I error effect sizes for research question three, LR controlled Type I error best only six times, in six cases LR and MH had equal Type I error rates, and LR prevailed with respect to Type I error control in the remaining 44 comparisons. Therefore, for the first three research questions which display Type I error data, MH has the best Type I error control the majority of the time; that is, in 44 out of 56 cases.

Research question four addresses Type II error concerns by comparing power percentages for the seven included studies (de Ayala et al., 2002; Herrera & Gomez, 2008; Kim & Oshima, 2012; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Vaughn & Wang, 2010) that displayed power data. When overall power data for these seven studies were averaged across all conditions LR had higher power rates indicating better performance for all studies except Herrera and Gomez (2008). When power data were separated according to uniform or nonuniform DIF, MH outperformed LR three to one. However, for the six studies with nonuniform DIF data, LR performed best five to one. In other words MH controlled Type I error best, and for the condition of uniform DIF, had the best power rates. However, LR controlled Type II error best by displaying the highest power rates for the nonuniform DIF condition.

Research Question 1

The first research question compares Type I error rates between MH and LR statistical methods. Since the Type I error rate is the number of unbiased items erroneously identified as DIF-containing divided by the total number of non-DIF items on the test, the statistical method displaying the lowest Type I error rate is the more desirable method for practitioners when considering Type I error alone. This is the case because using the method with the lower Type I error rate will result in fewer non-DIF test items being discarded unnecessarily.

Type I error rates were compared for LR and MH for the substantive study characteristics impact; that is, ability distribution, and sample size. Two levels of impact, equal and unequal, were examined and six levels of sample size, three levels of equal sample size and three levels of unequal sample size. Since average data by condition was shown only in aggregate form for Narayanan and Swaminathan (1996), Type I error values for impact and sample size are shown separately for that study. The three levels of equal sample size for the impact equal to zero condition and for the impact equal to one condition are shown in Figures 8 and 9, respectively. In the 11 sections below, the findings are discussed by sample size and impact.

Small, equal sample size and impact equal to zero. For the condition of small, equal sample size and impact equal to zero; that is, equal ability distributions, MH displayed lower Type I error rates for each of the five studies, except Vaughn and Wang (2010) which displayed equal Type I error rates for both MH and LR. For Vaughn and Wang (2010) Type I error rates for both MH and LR exceeded the nominal .05 level with Type I error rates of .090. The only other study for which the Type I error rate exceed the nominal .05 rate was DeMars (2009) and then only for LR with .055, just barely over the nominal level. Type I error rates for the small, equal sample size and impact equal to zero condition are shown in Figure 8.

Medium, equal sample size and impact equal to zero. For the equal, medium sample size and impact equal to zero condition, MH displayed lower Type I error rates than LR for five of the seven studies. MH Type I error rates exceed .05 for only one study, Herrera and Gomez (2008). Vaughn and Wang (2010) had equal rates of .020. LR Type I error rates exceeded the nominal .05 rate for five of the seven studies. Type I error rates for the medium, equal sample size, and impact equal to zero conditions are shown in Figure 8.

Large, equal sample size and impact equal to zero. For the large, equal sample size and impact equal to zero condition, two studies had equal Type I error rates for MH and LR: Vaughn and Wang (2010), 0; and Güler and Penfield (2009), .045. LR Type I error rates were lower than MH for two of three remaining studies with MH having the lowest rate for only one study. Therefore, LR controlled Type I error best for this particular condition. Type I error rates for the large, equal sample size, and impact equal to zero condition are shown in Figure 8.

Small, equal sample size and impact equal to one. The trend of MH exhibiting lower Type I error rates continued for the small, equal sample size and impact equal to one condition. Here two studies were compared (DeMars, 2009; Güler & Penfield, 2009) and MH was lower than LR for both. LR exceed the nominal Type I error rate for DeMars (2009), .068, and Güler and Penfield (2009). Type I error rates for the small, equal sample size and impact equal to one condition are shown in Figure 9.

Medium, equal sample size and impact equal to one. The only study to satisfy this condition was Narayanan and Swaminathan (1996). Again MH displayed lower Type I error rates overall; of the four Type I error rates only one, the sample size condition for MH, was below the .05 nominal rate. Type I error rates for the medium, equal sample size and impact equal to one condition are shown in Figure 9.

Large, equal sample size and impact equal to one. Three studies were compared for this condition, and though MH exhibited lower Type I error rates, the Type I error rates for DeMars (2009) were surprisingly large, MH of .097 and LR of .105. The difference between DeMars (2009), Güler and Penfield (2009) and Li, Brooks and Johanson (2012) was that the latter two studies had equal sample sizes of 1,000 while for DeMars (2009) values of 1,000 and

2,000 were averaged for the large range of equal sample size shown in Appendix X. Type I error rates for the large, equal sample size and impact equal to one condition are shown in Figure 9.

Unequal, small/medium sample size and impact of zero. Of the three studies compared (Herrera & Gomez, 2008; Narayanan & Swaminathan, 1996; Vaughn & Wang, 2010) all except Vaughn and Wang (2010) had Type I error rates above the .05 nominal rate for LR. On the other hand MH displayed Type I error rate values less than .05 except for Herrera and Gomez (2008), .055, and Vaughn and Wang (2010), .06. With reference to the other values Vaughn and Wang (2010) rates were MH, .06, and LR, .05. Type I error rates for the small/medium, unequal sample size and impact equal to zero condition are shown in Figure 10.

Unequal, small/large sample size and impact of zero. The trend of MH exhibiting lower Type I error rates continued for this condition. For Herrera and Gomez (2008) the Type I error rate was equal to .05, but it fell below the nominal value for the remaining two studies. For Narayanan and Swaminathan (1996) unequal sample size caused more Type I error rate inflation than impact equal to zero. Vaughn and Wang (2010) showed equal rates of .040 for MH and LR. Type I error rates for the small/large, unequal sample size and impact equal to zero condition are shown in Figure 10.

Unequal, medium/large, sample size and impact of zero. For all studies except Vaughn and Wang (2010), which had Type I error rates of .0100 for MH and LR, the Type I error rate for LR exceeded the nominal rate of .05. Herrera and Gomez (2008) joined Vaughn and Wang (2010) with equal Type I error rates, .060 and .010, respectively, for MH and LR. The largest Type I error rate for unequal sample size and impact of zero was that of Narayanan and Swaminathan (1996), with a value of .089. For all ranges of unequal sample size and impact of zero, Narayanan and Swaminathan (1996) displayed the highest Type I error rate for sample size

and LR, which seemed to indicate that that sample size had a greater effect on Type I error inflation than impact equal to zero, at least for that study. Type I error rates for the medium/large, unequal sample size and impact equal to zero condition are shown in Figure 10.

Unequal, small/medium, small/large, and medium/large sample size conditions for impact of one. For the conditions of small/medium and small/large sample size and impact of one, the only study with comparable data was Narayanan and Swaminathan (1996). Since the impact of one condition for Narayanan and Swaminathan (1996) was presented as an average in the article, Type I error rates are equal for impact in the unequal small/medium as well as small/large conditions. Though Type I error rates for MH and LR both exceeded the .05 nominal level, with the exception of MH for the sample size condition of Narayanan and Swaminathan (1996) for small/medium and small/large, MH controlled Type I error better across all three sample size conditions for impact equal to one for both compared studies. De Ayala et al. (2002) only simulated the medium/large sample size condition (500/2,500), for that condition MH performed better, 0.067, than LR, 0.076. Type I error rates for all unequal sample size and impact equal to one condition are shown in Figure 11.

Type I error rates averaged across sample size and impact. In addition to examining the effects of MH and LR on the Type I error for specific conditions by studies, comparisons were made of Type I error rates across impact by impact of zero and one as well as comparing impact of zero and impact of one divided by the six sample size conditions. When averaged across all sample size conditions, MH Type I error rates fell below the nominal .05 rate, while LR rates exceeded .05. Average Type I error rates across impact and sample size are shown in figures 13 and 14.

Separating Type I error by the six sample size conditions showed that overall for impact equal to zero and equal sample size MH had Type I error rates below the nominal .05 rate, except for the large, equal sample size condition, with MH once again displaying a lower Type I error rate in each instance. For unequal sample size and impact equal to zero, MH was superior, falling below the nominal rate for each condition, while LR values for Type I error were above MH for each unequal sample size condition.

The impact equal to one condition resulted in Type I error inflation for LR across all sample size conditions, and while MH had lower Type I error rates than LR for each condition, MH Type I error rates were below the nominal rate only for the equal small, medium and large conditions as well as the small/medium unequal condition. MH Type I error rate exceeded the nominal rate for unequal small/large and medium/large conditions.

Research Question 2

The second research question compared the Type I error rate of included studies to the nominal, or accepted, Type I error rate of .05. Using .05 as the accepted Type I error rate means that it is acceptable for a Type I error to occur for five percent of non-DIF items on a test. In Figures 14 through 17 the deviation scores from Type I error rate are depicted visually. In the following examples of calculations of deviation scores the Type I error rate is the first number in the equation and the nominal Type I error rate of .05 is the second number. Studies with Type I error rates of .05 are shown on the x-axis with values of zero (e.g., .05-.05 = 0), Type I error rates above .05 are shown above the x-axis as positive values (e.g., 0.06 – .05 = 0.01), and studies with Type I error rates below the .05 level have bars extending below the x-axis and are depicted with negative values (e.g., 0.04 – .05 = -0.01). Since lower Type I error rates are more desirable, because they indicate that at most five percent of unbiased test items have been

removed from the test, the lowest values; that is, the largest negative values or the longest bars below the x-axis show the studies that controlled Type I error best.

Type I error rate deviations from .05 for equal sample size and impact of zero. For Vaughn and Wang (2010) MH and LR shared deviation values for all levels of sample size. For the small, equal sample size condition the Type I error rate deviation of .0400 indicates Type I error rates above the nominal level, for the equal, medium condition, the negative values of -.300 for both methods indicate rates below the nominal rate, and for large sample size deviation values of -.050 meant Type I error rates were less than the nominal rate of .05.

For the condition of small, equal sample size MH outperformed LR with the exception of Vaughn and Wang (2010) discussed above. For DeMars (2009) LR was above the nominal rate with a deviation score of .005. For the remaining studies, except Swaminathan and Rogers (1990), LR deviation scores were negative indicating they fell below the nominal Type I error rate, though MH values were more negative; that is, extending further below the x-axis, than LR, giving MH the best control of Type I error inflation for the small, equal sample size condition with impact of zero.

MH's trend displaying lower Type I error deviation scores than LR continued for the equal, medium sample size condition. Here, all Type I error deviation scores for MH are less than zero, except for Herrera and Gomez (2008) with a value of zero. Aside from the equal deviation scores (Vaughn & Wang, 2010), all values for MH are lower than LR. Deviation scores for LR were above the nominal rate for all included studies but two (Kim & Oshima, 2012; Vaughn & Wang, 2010) in this condition. For Narayanan and Swaminathan (1996) and Swaminathan and Rogers (1990) the condition of sample size for LR proved most difficult for controlling Type I error inflation.

For the equal, large sample size condition with impact of zero, LR deviation scores were above zero for three studies. While DeMars (2009) showed only slight inflation, Herrera and Gomez (2008) displayed deviation scores of .0300 and .0200, for MH and LR, respectively, which were sizable for that condition. Equal deviation scores for MH and LR were exhibited by Vaughn and Wang (2010) and Güler and Penfield (2009).

Type I error deviations from .05 for equal sample size and impact of one. Though MH outperformed LR for all sample size conditions, LR had negative deviation scores for Güler and Penfield (2009) in the large sample size condition. Across all conditions, MH exceeded .05 for only two studies (Narayanan & Swaminathan, 1996; Güler & Penfield, 2009). DeMars (2009) showed uncharacteristically large deviation scores, .047 and .055, for the large sample size condition for MH and LR respectively. Type I error deviation scores for the equal sample size and impact equal to one condition are shown in Figure 14.

Type I error deviations from .05 for unequal sample size and impact of zero. Vaughn and Wang (2010) had equal deviation scores of zero for the small/large condition as well as the medium/large condition. For Herrera and Gomez (2008) equal scores of .010 for MH and LR for the medium/large condition were indicative of Type I error inflation. MH deviation scores were above zero in three other instances: the small/medium condition for Herrera and Gomez (2008) with a value of .0050, the small/medium condition for Vaughn and Wang (2010) with a value of .0100, and the medium/large condition for sample size for Narayanan and Swaminathan (1996) with a value of .0060. MH had a zero deviation score for Herrera and Gomez (2008) for the small/large condition; for the remainder of the studies deviation scores for MH fell below zero. In contrast, LR displayed values at or below zero for Vaughn and Wang (2010) for each condition, but deviation scores exceeded zero for rest of the studies in this condition meaning

that Type I error was not controlled with respect to the .05 nominal value. Deviation scores for all studies in the unequal sample size and impact equal to zero condition are shown in Figure 16.

Type I error deviations from .05 for unequal sample size and impact of one. Two studies simulated the impact equal to one condition for unequal sample size. De Ayala et al. (2002) only used the medium/large sample size condition, and deviation scores for MH and LR exceeded zero, though MH had a lower deviation score. Since data from Narayanan and Swaminathan (1996) was averaged by condition, sample size and impact conditions are graphed separately for that study. Impact data, which are the same for each sample size condition, indicate that while MH controlled Type I error best, both studies have deviation score above zero of, .005 and .048, respectively for MH and LR. For small/medium and small/large sample size conditions MH showed negative deviation scores in three out of five cases, while LR had all positive ones in keeping with the general trend indicating superior performance overall for MH over LR in Type I error control. For the medium/large sample size condition, neither MH nor LR had negative deviation scores, however, the lower score of MH indicates its' superiority over LR in the case of impact equal to one and medium/large sample size. Figure 17 provides a visual comparison of the three levels of unequal sample size for impact equal to one.

Research Question 3

Answering the third research question required the calculation of a statistic that could be used to compare all included studies on a common scale. Effect size was chosen as that statistic. For this study a proportion –based effect size, Type I error effect size, was calculated. In Figures 18 through 22 Type I error effect size is compared by condition and across studies for the two levels of impact and six levels of sample size. The first step in calculating Type I error effect size was taking the difference between MH and LR Type I error rates for each observation of data

provided in the included articles. Since LR was subtracted from MH; that is, MH-LR, negative effect size values indicated that MH had the lower Type I error rate and therefore controlled Type I error best, while positive Type I error effect size values indicated the converse, better Type I error inflation control via LR. A zero value for Type I error effect size simply means that the Type I error for the two methods, MH and LR, were the same, but does not provide any other information.

Type I Error effect size of studies with equal sample size and impact equal to zero. Vaughn and Wang (2010) had equal values for MH and LR for all equal sample sizes; the only other instance of equal Type I error for MH and LR was for the large, equal sample size condition (Güler & Penfield, 2009). Aside from instances of equal Type I error, MH outperformed LR for all studies except Herrera and Gomez (2008) in the large sample size condition and Kim and Oshima (2012) in the medium and large sample size conditions.

Type I Error effect size of studies with equal sample size and impact equal to one. Since all effect sizes, except Li et al. (2012) which was zero, were negative, MH was favored over LR for control of Type I error inflation for this condition. The lowest effect sizes occurred in the medium condition for the separate conditions of sample size, -.037, and impact, -.043 (Narayanan & Swaminathan, 1996).

Type I Error effect size of studies with unequal sample size & impact equal to zero. Though studies differed from the unequal to the equal sample size conditions with impact of zero, the overall results were similar. For instance, for the unequal, small/medium sample size group, Vaughn and Wang (2010) was the only study for which LR outperformed MH. Vaughn and Wang (2010) had effect sizes of zero for the small/large and medium/large unequal sample size conditions, while Herrera and Gomez (2008) showed and effect size of zero for the

medium/large condition. Narayanan and Swaminathan (1996) displayed the lowest effect sizes, and hence the best Type I error control, across studies in this condition. Of the twelve effect sizes calculated for the unequal sample size condition and impact equal to zero, MH was superior in eight out of twelve cases, with three ties between MH and LR, and LR being superior for Type I error control in one instance. Figure 20 displays graphically the results discussed for this condition.

Type I Error effect size of studies with unequal sample size and impact equal to one. MH keeps the lead over LR for this condition summarizing results from Narayanan and Swaminathan (1996) and de Ayala et al. (2002). For Narayanan and Swaminathan, impact equal to one produces lower effect sizes than sample size and the trend is reversed when impact equals zero across all conditions. These results can be seen in Figure 21.

Type I Error effect size averaged across sample size conditions. Averaging Type I error effect size across sample size produced two results, -.013 for the impact equal to zero condition, and -.025 for the impact equal to one condition. The fact that both effect sizes were negative meant that each condition favored MH, with the impact equal to one condition falling further below the .05 level. Data discussed here are shown pictorially in Figure 22.

Type I Error effect size averaged across impact. When effect sizes were averaged across impact and displayed according to sample size, MH prevailed. For equal sample size, the medium condition controlled Type I error best followed by small and then large, equal sample sizes. For unequal sample sizes, the small/large condition showed the best Type I error inflation control, followed by medium/large and small/medium conditions.

Average Type I error effect size. Showing the average effect sizes for sample size separated by impact changed the landscape. Once again, negative effect sizes demonstrated the

efficacy of MH for Type I error control. For impact equal to zero, the medium condition showed the lowest effect size, -.023, while the large condition was the only condition for which LR surpassed MH with an effect size of .002. For impact equal to one all sample sizes equal and unequal favored MH, with effect sizes for unequal sample size being the smallest, that is controlling Type I error best, for impact equal to one. Effect sizes averaged across sample sizes and divided by impact conditions are shown in Figure 22.

Research Question 4

Though inclusion criteria for this study did not specify the presence of power data, four of the included studies contained power data. Therefore, research question four evolved with this study. Information about power is also important for test development since power is the percentage of DIF-containing items that were correctly identified. Subtracting the power proportion from one provides Type II error which is the proportion of DIF-containing test items that were erroneously left on a test. Since the data in the graphs is shown as power, the higher numbers indicate the statistical method, LR or MH, which performs better with regard to correct identification of DIF-containing items. Figures 23 and 24 compare power for MH and LR.

Since DIF effect data accompanied power for only three of the seven studies displaying power data, power data was presented as an average for each of the seven studies displaying power data. Four studies simulated the condition of uniform DIF, and six studies simulated nonuniform DIF. Therefore, in addition to displaying overall average power data in Figure 23, average power data broken by down uniform and nonuniform DIF conditions are shown in Figure 24. For the overall average power comparison LR displayed higher rates for correct identification of DIF items, which is the desired result. When power data were split according to uniform versus nonuniform DIF, MH performed slightly better than LR for two out of four

studies. For Herrera and Gomez (2008) LR showed a mysteriously low average power rate. Since the data point seemed incongruent with respect to the other data points, I contacted the authors. At this time I have not received a reply concerning the .17 power rates for LR produced by Herrera and Gomez (2008). The third study simulating uniform DIF displayed a .82 power rate for LR and .68 for MH, placing LR ahead of MH just one in for times for the condition of uniform DIF. In the case of nonuniform DIF LR power rates exceeded those of MH by a comfortable margin in four out of six cases. For de Ayala et al. (2002) MH and LR tied and for Vaughn and Wang (2010) MH was ahead only by a nose with a power rate of .50 to the .49 rate exhibited by LR for that study. The marked difference in correct identification, or power rates, for uniform versus nonuniform DIF between MH and LR is no surprise since the phenomenon was demonstrated by Rogers and Swaminathan (1993). Therefore for control of Type I error and higher power rates for correct identification with uniform DIF, MH is recommended, while for the highest power rates with regard to nonuniform DIF, LR is recommended.

Type I or Type II Error, That is the Question

Type I error is the focus of this paper because Type I error was one of the inclusion criteria. Inclusion criteria specifying the presence of Type I error rate data in usable form placed Type I error rate comparisons in the forefront of this paper. However, recognizing the importance of Type II error, MH and LR power percentages were compared in the fourth research question for the four studies with power rates. A 2 x 2 table has been used (Gill 1978) to compare and contrast the two types of error, as shown in Figures 23 and 24. The real priority of Type I versus Type II error control is a situational one. In most cases, the priority is dictated by cost as stated by Smith (2012), "the question of whether to choose a low Type I error rate or a low Type II error rate is actually asking whether it is more costly to allow false positive or false

negative results." The formula for a loss function pertaining to Type I and II error can be represented by,

Loss = *P*(Type I error) x Loss for Type I error + *P*(Type II error) x Loss for Type II error (5.1).

For example, this means that the probability of a Type I error multiplied by the financial loss caused by the Type I error added to the probability of Type II error multiplied by the financial loss caused by a Type II error gives the total loss caused by Type I and II errors combined. It follows then that calculating the loss for a Type I error and then the loss for a Type II error could provide guidance for deciding which type of error is most important for a given situation.

Type I error. Type I error or alpha (α), also called a sin of commission (Light, 1991), indicates a false positive situation in which a non-biased item or items are flagged as DIF-containing and removed from a test unnecessarily. The costs related to such an error would be the development costs of having to create a new item. If time constraints did allow for the development of an item to replace the biased one, then the test would be shorter which could decrease the reliability of

the test. In addition to decreasing the reliability, earning a passing score could be more difficult for examinees taking a shorter test since each item would carry more weight in lost points.

Type II error. Type II error or beta (β), sometimes referred to as a sin of omission (Hansen, 2005), is a false negative situation which occurs when a DIF-containing item is not flagged and thus remaining on the test. This type of error could potentially incur greater costs if the biased item was discovered and the equity of the test across different groups was questioned. In that case, the disadvantaged group might have grounds to refute the validity of test results

Table 6.

Type I Error and Type II Error Decision versus State of Nature for DIF

Decision	State of Nature: H_o is true (no DIF)	State of Nature: H_o is false (DIF)
Accept H_o	Correct Decision (CD) $P(CD) = 1 - \alpha$	Type II error $P(\text{Type II error}) = \beta$
Reject H_o	Type I error $P(\text{Type I error}) = \alpha$	Correct Decision (CD) $P(CD) = 1 - \beta$

depending in part, on the item's contribution to the test score and subsequently decisions based on the test.

To close the discussion on the selection of an appropriate statistical method with respect Type I and Type II error Keselman, Games, and Rogan (1980) provide this perspective:

> A perpetual dilemma in statistical inference is that there are two types of error, and (other things held constant) reducing the risk of one increases the risk of the other. In some cases, the relative importance of these two types of error may be guided by the nature of the research. The decision is not a mathematical judgment but rather a subjective one, and "every man should get to pick his own error rates" (Miller, 1966, p.33, as cited in Keselman, Games & Rogan, 1980).

That is to say that with regard to the choice of the appropriate statistical method, LR or MH, for identification of DIF each situation calls for independent examination and practitioners should chose the method that does the best job of minimizing loss for their particular case. The overall recommendation is for practitioners whose preference is control of Type I error to use MH and for practitioners whose top priority is power to use LR for identification of DIF-containing items.

Limitations and Future Research

Though over a dozen statistical methods exist for evaluation of DIF (Appendix A), there are currently approximately 10 methods in use. While this study has collated currently existing

data for MH and LR statistical methods for DIF detection, a limitation of this study is the lack of comparative data for all conditions presented in the included studies. DeMars (2009) and Li, Brooks and Johanson (2012) varied the impact or ability difference between the commonly found values of zero and one. Additional values for impact were provided by three studies (DeMars, 2009; Li, Brooks & Johanson, 2012; Vaughn & Wang, 2010) but these values were unique to each study and thus, lacked a basis for comparison. While DeMars (2009) and de Ayala et al. (2002), experimented with large sample values of 2,000 and 2,500 respectively, they were the only ones to do so. Li, Brooks and Johanson (2012) was the sole study to simulate replications of 10,000; Vaughn and Wang (2010) with replications of 1,000, was the only study with number of replications under 10,000 that exceeded 300.

One limitation was the inability to incorporate the calculation of the arcsin transformed effect size, which could have added not only an additional value for effect size but also a potentially different value. A second limitation was the variety of simulation parameters and values of those parameters across studies. A comparison of study characteristics and parameters is provided in Appendix J. A recommendation pertaining to parameter selection would be to run a simulation study with parameters in common with published studies. Then run a simulation study with experimental parameters. See Tran (2011) for an example. Following this methodology would provide a solid basis for comparison of existing studies. Research extensions could include a more extensive exploration of the literature and an exploration of real-data studies as well as simulation studies. Also, additional effect size measures could be calculated using means and standard deviations in addition to Type I error data.

In conclusion, this meta-analysis quantitatively summarized 10 published studies to provide findings regarding the Type I error control of MH and LR statistical methods for DIF

detection. Summative data were presented as Type I error rates, Type I error deviation scores, and Type I error effect sizes. Using the meta-analysis software package, MetaAnalyst, treatment effects and confidence intervals were calculated for each study. Power data were presented for four studies on one level of impact and two levels of equal sample size. Finally, study characteristics and study parameters were summarized in an effort to organize the current research and with the positive side effect of creating a succinct table for easy access by authors of future simulation studies who wish to create a segue between past and future studies or who wish to cover new ground.

References

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement, 18*(4), 329–342.

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI. Retrieved from the ERIC database. (ED 069 686)

Bielinski, J. D., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*(3), 455–476. doi:10.2307/1163444

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113–141.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Bradbury, A. (2011). Rethinking assessment and inequality: the production of disparities in attainment in early years education. *Journal of Education Policy, 26*(5), 655–676.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*(2), 101–125.

Camili, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*(1), 12–37.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaptation-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*(2), 169–199.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's chi square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39–52.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10*(4), 237–255.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*(10), 1067–1077.

Cooper, H. (2004). *Research synthesis and meta-analysis: A step by step approach* (4th ed.). Los Angeles, CA: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Coulter-Kern, M., Coulter-Kern, R., Howard, G. S., Hill, T. L., Maxwell, S. E., Baptista, T. M., Coelho, C. (2009). What's wrong with research literatures? And how to make them right. *Review of General Psychology, 13*(2), 146–166.

Curlette, W. L., & Cannella, K. S. (1985). Going beyond the narrative summarization of research findings: The meta-analysis approach. *Research in Nursing and Health, 8*, 293–301.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Education Measurement*, 8(2), 71-82.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

*de Ayala, R. J., Kim, S., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*(3–4), 243–276.

DeVellis, R. F. (2011). Scale development: theory and applications. *Sage,* Thousand Oaks, CA.

*DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34*(2), 149–170.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszsel method. *Applied Measurement in Education, 2*(3), 217–233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23,* 355–368.

Egger, M., Davey Smith, G. Schneider, M., & Minder, C. E. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315*(7129), 629–34.

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113–148.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Ferguson, C. J., & Brannick, M. T. (2011). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods 17*(1), 120–128.

Fidalgo, A. M., Ferreres, D., & Muniz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *The Journal of Experimental Education, 73*, 23–39.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muniz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, 75*(4), 293–314.

Finch, W. H. (2011). The impact of missing data on the detection of non-uniform differential item functioning. *Educational and Psychological Measurement, 71*(4), 663–683.

Finch, W. H., & French, B. F. (2008). Anomalous Type I error rates for identifying one type of DIF in the presence of another. *Educational and Psychological Measurement, 68*, 742–759.

French, B. F., & Finch, W. H. (2013). Extensions of Mantel-Haenszel for multilevel DIF detection. *Educational and Psychological Measurement, 73*(4), 648–671.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373–393.

Furlow, C. F., Ross, R., & Gagné, P. (2009). The impact of multidimensionality of on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*(6), 441–464.

Garrett, P. L. (2009). *A Monte Carlo study investigating missing data, differential item functioning, and effect size* (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Garvey, W. D., & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist*, *26*(4), 349–362.

Gill, J. L. (1978). Design and analysis of experiments in the animal and medical sciences. Ames, IA: The Iowa State University Press.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Education Researcher*, *5*, 3–8.

Glass, G. V. (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351–379.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Newbury Park, CA: Sage.

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J. (2009). Efficacy of effect size measures in logistic regression: an application for detecting DIF. *Methodology*, *5*(1), 18–25.

Gleser, L. J. & Olkin, I. (2009). Special Statistical Issues and Problems. In Cooper, H., Hedges, L. V., & Valentine, J. C. Editors, *The handbook of research synthesis and meta-analysis* (2nd ed.) p. 364. New York, NY: Russell Sage Foundation.

Gonzalez-Roma, V., Hernandez, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*, 29–53.

Greatorex, J., & Bell, J. (2004) Does the gender of examiners influence their marking? *Research in Education, 71*, 25–36.

Guilera, G., Gómez-Benito, J., & Hidalgo, M. (2010). Citation analysis in research on differential item functioning. *Quality & Quantity*, *44*(6), 1249–1255.

*Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and non-uniform DIF. *Journal of Educational Measurement, 46*(3), 314–329.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Journal of Research in Education*, *2*(4), 313–334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hansen, D. M. (2005). A sin of omission: database transactions. *Journal of Computing Sciences in Colleges*, 21, 1, 225-230.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35–41.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

*Herrera, A. N., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality and Quantity, 42,* 739-755.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Hidalgo, M. D., & Gomez, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: a comparison for polytomous items. *Quality and Quantity, 40,* 805–823.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903–915.

Higgins, J. P. T. (2008). Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology, 37*, 1158-1160.

Higgins, J., & Green, S. (2008). *Cochrane handbook for systematic review of interventions: Cochrane book series*. Chichester, UK: Wiley.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Holland, P. W. (1985, October). *On the study of differential item performance without IRT.* Paper presented at the meeting of the Military Testing Association, San Diego, CA.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*(2), 227–240.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology, 37*(1), 47–61.

Keselman, H. J., Games, P.A., Rogan, J.C. (1980). Type I and Type II errors in simultaneous and two-stage comparison procedures. *American Psychological Association, 88,*2, 356-358.

Kim, J. (2006). Using the distractor categories of multiple-choice items to improve IRT linking. *Journal of Educational Measurement, 43*, 193–213.

Kim, J. (2010). *Controlling Type I error rate in evaluation differential item functioning for four DIF methods: use of three procedures for adjustment of multiple item testing.* (Doctoral Dissertation, Georgia State University). Retrieved from http://digitalarchive.gsu.edu/eps_diss/67

*Kim, J., & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458–470.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93–116.

Kirshner, B. & Guyatt, G. (1985). A methodological framework for assessing health indicies. *Journal of Chronic Diseases*, 38(1), 27-36.

Kwak, N., Nohoon, Davison, M. L., & Davenport, E. C., Jr. (1997, April). *An unsigned Mantel-Haenszel statistic for detecting uniform and non-uniform DIF*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T.C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. The *New England Journal of Medicine, 327*, 248–254.

*Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*,5, 847-861.

Light, D.W. (1991). Effectiveness and efficiency under competition: the Cochrane test. *British Medical Journal*, 303, 6812, 1253-54.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: procedures for resolving contradictions among different research studies. *Harvard Education Review, 41*(4), 429–471.

Linn, R. L., & Drasgow, F. (1987). Implications of the golden rule settlement for test construction. *Educational Measurement: Issues and Practice, 6*(2), 13–17.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford.

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent Type I error in differential item functioning. *Educational and Psychological Measurement, 72*(2), 291–311.

Magis, D., & Facon, B. (2012). Item purification does not always improve DIF detection: a counter example with Angoff's delta plot. *Educational and Psychological Measurement, 73*(2), 293–311.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

MetaAnalyst. Meta-analysis software package. Retrieved from http://sites.tufts.edu/statisticalcomputingmatters/ April 17, 2013.

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on Type I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Education Measurement*, 42(2), 101–131.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131–144.

Meade, A. (2010). A taxonomy of effect size measures for the differential item functioning of items and scales. *Journal of Applied Psychology*, *4*, 728–743.

*Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20, 257–274.

Nankervis, B. (2011). Gender inequities in university admission due to the differential validity of the SAT. *Journal of College Admissions, 213,* 24–30.

Oosterhof, A. C., Atash, M. N., & Lassiter, K. L. (1984). Facilitating the identification of item bias through the use of delta plots. *Educational and Psychological Measurement*, *44*(3), 619–627.

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning*. Second edition. Sage, Los Angeles.

Penfield, R. D., Alvarez, K., & Lee, O. (2010). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: an illustration. *Applied Measurement in Education*, 22, 61-78.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: a comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*(3), 235–259

Price, D. J. (1965). Networks of scientific papers: the pattern of bibliographic references indicates the nature of the science research front. *Science*, *149*(3683), 510–515.

Newman, D. A., Hanges, P. J., & Outtz, J. L. (2007). Racial groups and test fairness, considering history and construct validity. *American Psychologist, 62*(9), 1082–1083.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential function of items and tests. *Applied Psychological Measurement, 19,* 353–368.

Rivas, G. E. L., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265.

Robitzsch, A., & Rupp, A. A. (2009). The impact of missing data on the detection of differential item functioning: the case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*(1), 18–34.

*Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105–116.

Rosenthal, R. (1994). Parametric measures of effect size. In Cooper, H., & Hedges, L. V. Editors, *The handbook of research synthesis* (p. 237). New York, NY: Russell Sage Foundation.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., the Quality of Life Cross-Cultural Meta-Analysis Group. (2010). Differential item functioning analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes, 8*(81), 1–9.

Sipe, T. A., & Curlette, W. L. (1997). A meta-synthesis of factors relating to educational achievement: A methodological approach to summarizing and synthesizing meta-analyses. *International Journal of Education Research*, *25*, 583–698.

Smith, C.J. (2012). Type I and Type II errors: what are they and why do they matter? *Phlebology*, *27*, 4, 199-200.

Sanchez-Mecca, J., & Marin-Martinez, F. (1997). Homogeneity tests in meta-analysis: a Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity*, 31, 395–399.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician*, *40*(2), 106–108.

Spray, J., & Miller, T. (1994) *Identifying non-uniform DIF in polytomously scored test items. ACT research report series 94–1*. Iowa City, IA: American College Testing Programs.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*(4), 313–350.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. New York, NY: Wiley.

*Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Thalheimer, W., & Cook, S. (2002). *How to calculate effect sizes from published research: a simplified methodology.* Available online: http://www.docstoc .com/docs/47860289/How-to-calculate-effect-sizes-from-published-research-A-simplified-methodology

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics*, *21*(2), 110–130.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001. Technical report*. Minneapolis, MN: University of Minnesota.

Thurman, C. (2009). *A Monte Carlo study investigating the influence of item discrimination, category intersection parameters, and differential item functioning patterns on the detection of differential item functioning in polytomous items*. (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Tran, D. M. (2011). *Robustness of Two Formulas to Correct Pearson Correlation for Restriction of Range*. (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA. http://scholarworks.gsu.edu/eps_diss/84

Traver, D. F., Aliste, A. M. P., & Muniz, J. (2000). Detection of non-uniform DIF: Mantel-Haenszel and logistic regression methods. *Psicothema Revista De Psicologia, 12,* 2.

*Vaughn, B. K., & Wang, Q. (2010). DIF trees: using classification trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6), 941–952.

van de Vijver, F. & Hambleton, R. K. (1996). Translating tests: some practical guidelines. *European Psychologist*, 1(2), 89-99.

Wang, C. M., & Bushman, B. J. (1999). *Integrating results through meta-analytic review using SAS software.* Cary, NC: SAS Institute.

Wang, W. (2004). Effects of anchor item methods on the detection of differential item functioning with the family of Rasch models. *The Journal of Experimental Education*, *72*, 221–261.

Wang, W., & Su, Y. (2004). Effects of average signed area between two items characteristic curves and test purification procedures on the DIF detection via Mantel-Haenszsel method. *Applied Measurement in Education, 17*(2), 113–114.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: problems and an example. *Journal of Educational Measurement, 32*(2), 163–178.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Education and Psychological Measurement*, 59(6), 910–227.

Wiberg, M. (2009). Differential item functioning in mastery tests: a comparison of three methods using real data. *International Journal of Testing, 9*, 41–59.

Wilkinson Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

Woods, C. M. (2009). Testing for differential item functioning with measures of partial association. *Applied Psychological Measurement, 33*, 538–554.

Woods, C. M., & Grimm, K. J. (2011). Testing for non-uniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339–361.

Zappe, S. E. (2007). *Response process validation of equivalent test forms: how qualitative data can support the construct validity of multiple test forms* (Unpublished doctoral thesis). Pennsylvania State University, University Park, PN.

Zhang, M. (2009). *Gender related differential item functioning in mathematics tests: a meta-analysis* (Unpublished master's thesis). Washington State University, Pullman, WA.

Zheng, Y., Gierl, M. J., & Cui, Y. (2007). *Using real data to compare DIF detection and effect size measure among Mantel-Haenszel, SIBTEST, and logistic regression procedures*. (Unpublished thesis). University of Alberta, Edmonton, CN.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., & Erickan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 2655–2666.

Zwick, R., Thayer, D., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321–344.

*Indicates included study

APPENDIXES

APPENDIX A

Methods for Detection of DIF

Method of DIF detection	Source	Advantages	Disadvantages
ANOVA-based methods	Whitmore & Shumacker (1999)		
Area between 2 Item Response Functions	Kim & Cohen, (1991); Raju, (1988, 1990); Rudner, Geston, & Knight (1980)		
Breslow-Day	Breslow & Day (1980)	powerful for detecting crossing-non-uniform DIF	not powerful for detecting uniform DIF
Delta Method	Angoff & Ford (1973)		
Graded Response Model	Samejima (1969)		
Graded Response Model-Differential Functioning of Item and Tests	Flowers, Oshima, & Raju (1999)		
Graded Response Model-Likelihood Ratio	Thissen, Steinberg & Gerrard (1986)		
IRT Logistic Regression	Thissen, Steinberg, & Gerrard (1986); Thissen et al. (1988)		
IRT Rasch Model	Rasch (1960)		
IRT Three-Parameter Logistic Model	Hambleton, Swaminathan, & Rogers (1991); Lord & Novick (1968)		
IRT Two-Parameter Logistic Model	Lord (1952)		
likelihood ratio test	Thissen, Steinberg, & Wainer (1988)		

Method of DIF detection	Source	Advantages	Disadvantages
Logistic analysis, Discriminant	Miller & Spray (1993)	polytomous data	
Logistic Regression	Swaminathan & Rogers (1990); Rogers & Swaminathan, (1993)	displays good power for detecting uniform and non-uniform DIF	1) inflated Type I error 2) requirement of iterative parameter estimation (makes it computationally expensive)
LR, Multinomial	French & Miller (1996); Miller & Spray (1993); Miller et al. (1992)	can handle polytomous data	1) requires large amounts of data manipulation 2) interpretation of results is difficult because many parameters have to be tested to statistical significance
Lord's chi-square	Lord (1980)		
Mantel Method	Mantel (1963)	can handle polytomous data	
MH Generalized	Somes (1986)	can handle polytomous data	
MH adapted			
MH common odds ratio	Holland & Thayer (1988); Dorans & Holland (1993); Mantel & Haenszel (1959)		
MH two-stage			
MH iterative			
MH with chi-square (no Adjustment)	Camilli & Shepard (1994); Mantel & Haenszel (1959)	1) most powerful unbiased test for Ho of no DIF, 2) does not require large sample size	completely ineffective at detecting crossing-non-uniform DIF
MH with chi-square Bonferoni Adjustment			

APPENDIX B

Methodological Study Characteristics

Generating Model and Item Parameters for Non-studied (non-DIF) Items

Study	Generating Model	Item Parameters
De Ayala et al. (2002)	2PL	*b* parameters were randomly generated from N[0,1] distribution
DeMars (2009)	3PL	M = 0, SD = 2 [-2, 2]
Güler & Penfield (2009)	3PL	Güler & Penfield, 2009, p. 322
Herrera & Gomez (2008)	3PL	p. 743
Kim & Oshima (2012)	3PL	p. 462
Li, Brooks & Johanson (2012)	2PL	*b* parameters were randomly sampled from a uniform distribution (-2.0 to 2.0) *a* parameters had fixed ranges (i.e., 0.2 to 2.0 or 1.2 to 2.0) with a uniform distribution for different analyses
Narayanan & Swaminathan (1996)	3PL	p. 264
Rogers & Swaminathan (1993)	2PL 3PL	p. 109 p. 110
Swaminathan & Rogers (1990)	3PL	uniform DIF: *b* parameters were varied to produce DIF *a* parameters were fixed non-uniform DIF: *b* parameters were fixed at 0 *a* parameters were varied
Vaughn & Wang (2010)	1PL	*b* = -1 (13 items) *b* = 0 (14 items) *b* = 1 (13 items)

APPENDIX C

Methodological Study Characteristics

DIF Magnitude and Nature of DIF

Study	DIF Magnitude	Nature of DIF
de Ayala et al. (2002)	Moderate (Δb = 0.3) High (Δb = 1.0)	Non-uniform
DeMars (2009)	random	random
Güler & Penfield (2009)	Δb = 0 Δb = 0 Moderate (Δb = 0.25) Moderate (Δb = 0.25)	Uniform Crossing Non-uniform Non-crossing
Herrera & Gomez (2008)	Moderate (Δb = -1 to 1)	Uniform Non-uniform Mixed
Kim & Oshima (2012)	Small (Δb = 0.3) Medium (Δb = 0.5) Large (Δb = 0.7)	Uniform
Li, Brooks & Johanson (2012)	Δb = 0	Uniform
Narayanan & Swaminathan (1996)	Low (b = -1.5) Medium (b = 0) High (b = 1.5)	Non-uniform
Rogers & Swaminathan (1993)	Low (b = -1.5) High (b = 1.5)	Uniform
Swaminathan & Rogers (1990)	Moderate (Δb = 0.48) High (Δb = 0.64)	Non-uniform Uniform
Vaughn & Wang (2010)	Small (0.43)* Medium (0.64)* Large (0.86)*	Uniform

*logit scale see: Vaughn, B. K. (2008). Better quality in assessments: consideration of contextual effects on item bias and differential item functioning. *Journal on School Educational Technology*, *4*(2), 29-39.

APPENDIX D

Methodological Study Characteristics

Discrimination and Difficulty Parameter Differences for Studied Items With Studied Item Placement

Study	Studied item placement	Discrimination (*a* parameter)	Difficulty (*b* parameter)
de Ayala et al. (2002)	No DIF	1.0	0
	1-3	1.0	0.3
	1-3	1.0	1.0
	1-6	1.0	0.3
	1-6	1.0	1.0
DeMars (2009)	random	1.2, 1.4, 1.6, 1.8, 2.0	random
Güler & Penfield (2009)	1 (no DIF)	1.0	0
	2	1.3	0
	3	1.6	0
	4	1.0	0.25
	5	1.3	0.25
	6	1.6	0.25
Herrera & Gomez (2008)	4, 24, 43, 73	$a > 0.7$	-1 to 1
	9, 49, 70, 82	$a > 0.7$	-1 to 1
	18, 42, 53, 83	$a > 0.7$	-1 to 1
Kim & Oshima (2012)	20-item test	20-item test	20-item test
	9	1	0.3
	10	1	0.5
	11	1	0.7
	40-item test	40-item test	40-item test
	17-18	1	0.3
	19-20	1	0.5
	21-22	1	0.7

Study	Studied item placement	Discrimination (*a* parameter)	Difficulty (*b* parameter)
Li, Brooks & Johanson (2012)	random	fixed in different ranges (i.e., 0.2 to 2.0 or 1.2 to 2.0)	randomly sampled from uniform distribution (-2.0 to 2.0)
Narayanan & Swaminathan	n.s.*	Low *b*	-1.5
		Medium *b*	0
		Medium *b*	0
		High *b*	1.5
Rogers & Swaminathan (1993)	1	0.6	-1.5
	2	1	0
	3	1.6	1.5
	4	0.6	1.5
	5	1.6	-1.5
Swaminathan & Rogers (1990)	n.s.*	1	0.48
		1	0.64
		0.6 & 0.8**	0
Vaughn & Wang (2010)	2, 24, 37	n.s.*	DIF effect = 0.43
	6, 11, 32		DIF effect = 0.64
	9, 21		DIF effect = 0.86

*n.s. = not specified

**area between curves

APPENDIX E

Methodological Study Characteristics

Number of Replications per Study

Study	Number of Replications
de Ayala et al. (2002)	50
DeMars (2009)	300 (20 item test) 150 (40 item test) 100 (60 item test)
Güler & Penfield (2009)	200
Herrera & Gomez (2008)	100
Kim & Oshima (2012)	100 100
Li, Brooks & Johanson (2012)	10,000
Narayanan & Swaminathan (1996)	100
Rogers & Swaminathan (1993)	100
Swaminathan & Rogers (1990)	20
Vaughn & Wang (2010)	1,000

APPENDIX F

Comparison of Statistical and IRT Methods for the Detection of DIF

	Logistic Regression	Mantel-Haenszel	IRT
Advantages	Can evaluate for uniform and non-uniform DIF simultaneously	Has established effect size measure (Roussos & Stout, 1996b)	Uses ICCs which provide good visuals for increased understanding
	Can be readily expanded to handle two or more ability estimates (Swaminathan & Rogers, 1990)	May be considered the "gold standard" in DIF detection. (Roussos & Stout, 1996b)	Properties of invariance and ability parameters insure that tests & items are developed independent of examinees
Disadvantages	Costs 3-4 times as much as MH (Swaminathan & Rogers, 1990)	Designed to detect uniform DIF. May not detect non-uniform DIF (Swaminathan & Rogers, 1990)	"Does take into account the continuous nature of ability when comparing the performance of groups of examinees" (Swaminathan & Rogers, 1990).
	Results in high false positives (Gomez)	Not designed to detect non-uniform DIF	Requires large sample size
Possible solutions	Use effect size measure (Gomez)	Use to analyze tests where identification of non-uniform DIF is not essential	Everett Smith has used 250 as sample size.
	Use purification procedures for matching (Gomez, French)	Good for limited budgets.	Conduct simulation studies.

APPENDIX G

Included Studies with Data Type and Location

Study	Data Type	Location of Data
de Ayala, Kim, S. H., Stapleton, & Dayton (2002)	Number of times each item was identified as exhibiting DIF	pp. 254, 256, 258, 260, 262
DeMars (2009)	Type I error rate	p. 161, excel spreadsheet provided by author via email
Güler & Penfield (2009)	Rejection rates	p. 323
Herrera & Gomez (2008)	Type I error rate	p. 748
Kim, J. & Oshima (2012)	Type I error rate Average power rate (DIF magnitude = .5)	p. 165
Li, Brooks & Johanson (2012)	Type I error rate	pp. 854, 857, 858
Narayanan & Swaminathan (1996)	Average Type I error rate Average Power rate (DIF effect sizes of .4 & .6 averaged)	p. 267 p. 267
Rogers & Swaminathan (1993)	Type I error rate	p. 112
Swaminathan & Rogers (1990)	Type I error rate	p. 367
Vaughn & Wang (2010)	Average Type I error rate Average Power rate (Low, .43, & Medium, .64, DIF effects averaged)	p. 948

APPENDIX H

Implementing the Comparison of Nested Models for LR

	Comparison of nested models	Nature of DIF
Phase 1	Total score of the subject on the test (TOT) is introduced into equation 1 (model 1)	
Phase 2 Test for uniform DIF	The group variable (GENDER) is added to the equation (model 2)	
	R^2 model 1 – R^2 model 2 = (Zumbo, 1999)	Variation attributable to group differences (uniform DIF)
	$\tau_2 = \beta_{01} - \beta_{02}$ Group difference = intercept of group 1 –intercept of group 2 (Swaminathan & Rogers, 1990)	If slopes are equal ($\beta_{11} = \beta_{12}$) and intercepts are not equal ($\beta_{01} \neq \beta_{02}$)we have uniform DIF
	$R^2\Delta = R^2_2 - R^2_1$ (Kanjee, 2007)	
Phase 3 Test for non-uniform DIF	The interaction between group and total score is fitted to the equation (TOT*GENDER)	
	$\tau_3 = \beta_{11} - \beta_{12}$ Interaction difference= Slope of group 1 – slope of group 2 (Swaminathan & Rogers, 1990)	If slopes are different ($B_{11} \neq \beta_{12}$), we infer non-uniform DIF, regardless of the intercepts
	R^2 model 3 – R^2 model 2 = (Zumbo, 1999)	Relevance of interaction term (non-uniform DIF)
	$R^2\Delta = R^2_3 - R^2_2$ (Kanjee, 2007)	
G^2 likelihood ratio	R^2 model 3 – R^2 model 1	Are the group and interaction variables statistically significant over the matching criteria?
Model 2	Represents uniform DIF	
Model 3	Represents uniform and non-uniform DIF simultaneously	
Overall DIF	$R^2\Delta = R^2_3 - R^2_1$ (Kanjee, 2007)	

Gómez-Benito, Hidalgo & Padilla, 2009, p. 19

APPENDIX I

Summary of the Logistic Regression Equation Variable Meanings Applied for DIF Understanding the Notation for Nested Models

Author/variable name			Meaning of variable
Zumbo (1999)	Swaminathan & Rogers (1990)		
TOT (total test score)	Θ	Ability Estimate	Is the measure of ability often reflected by the total test score
GENDER (male or female)	G	Grouping Variable	The group variable (reference or focal)
b_0	τ_o	Intercept	The intercept parameter
b_1	τ_1	Slope	Ability difference parameter
b_2	τ_2	Degree of non-uniform DIF (Kanjee, 2007)	Group difference in performance on item parameter
b_3	τ_3	Degree of uniform DIF (Kanjee, 2007)	Parameter representing interaction between group and ability

APPENDIX J

Final Coding Table

	Impact	**DIF Magnitude**		**DIF % & Test Length**			**Studied Item Parameters & Placement**			**Sample Size**
Study	Ability Difference	DIF Magnitude	Nature of DIF	DIF %	Test Length (Replications)	# of Studied Items	Studied Item Placement	Discrimination (*a* parameter)	Difficulty (*b* parameter)	Focal/ Reference
de Ayala el al. (2002)	1	Δb = 0.3 Δb = 1.0	non-uniform	0% 10% 20%	30 (50)	0 3 3 6 6	No DIF 1-3 1-3 1-6 1-6	1	0 0.3 1.0 0.3 1.0	500/2,500
DeMars (2009)	0 0.25 0.5 0.75 1	random	random	30% 15% 10%	20 (300) 40 (150) 60 (100)	6	random	1.2 1.4 1.6 1.8 2.0	random	250/250 1,000/1,000 2,000/2,000
Guler & Penfield (2009)	0 1	Δb = 0 Δb = 0.25 Δb = 0 Δb = 0.25	uniform crossing non-uniform non-crossing	10%	60 (200)	6	1 (no DIF) 2 3 4 5 6	1.0 1.3 1.6 1.0 1.3 1.6	0 0 0 0.25 0.25 0.25	300/300 300/1,000 1,000/1,000
Herrera & Gomez (2008)	0	Δb = -1 to 1	uniform non-uniform mixed	12%	100 (100)	12	4, 24, 43, 73 9, 49, 70, 82 18, 42, 53, 83	a > 0.7	-1 to 1	500/500 1,500/1,500 100-250/500 300-750/1,500

	Impact	DIF Magnitude & Nature of DIF		DIF % & Test Length			Studied Item Parameters & Placement			Sample Size
Study	Difference	Magnitude	DIF	DIF %	(Replications)	Studied	Placement	(*a* parameter)	(*b* parameter)	Reference
Kim & Oshima (2012)	0	Δb = 0.30 Δb = 0.5 Δb = 0.7	uniform	0.15	20 (100)	2	9 10 11	1	0.3 0.5 0.7	500/500 1,000/1,000
					40 (100)	4	17-18 19-20 21-22	1	0.3 0.5 0.7	
Li, Brooks & Johanson (2012)	1	Δb = 0	uniform	0%	50 (10,0000)	1*	random	fixed in different ranges (i.e., 0.2 to 2.0 or 1.2 to 2.0)	randomly sampled from uniform distribution (-2.0 to 2.0)	300/300 650/650 1,000/1,000
Narayanan & Swaminathan (1996)	0 1	Δb = 0	non-uniform	0% 10% 20%	40 (100)	16	n.s.*	high *a* > .47 low *a* < .5	low *b*= -1.5 medium *b*=0 medium *b*=0 high *b*=1.5	500/500 200/500 200/1,000 500/1,000
Rogers & Swaminathan (1993)	1	Δb = 0	uniform	0	40 (100)	5*	1 2 3 4 5	0.6 1 1.6 0.6 1.6	-1.5 0 1.5 1.5 -1.5	250/500
Swaminathan & Rogers (1990)	0	n.s.*	non-uniform uniform	20%	40 (20) 60 (20) 80 (20)	8 12 16	n.s.*	1 1 0.6 & 0.8**	0	250/500
Vaughn & Wang (2010)	0 0.5		uniform	20%	40 (1,000)	2, 6, 9, 11 11, 42, 32, 37	2, 24, 37 6, 11, 32 9, 21	n.s.*	DIF effect = 0.43 DIF effect = 0.64 DIF effect = 0.86	250/250 500/500 1,000/1,000 250/1,000 500/1,000

*not specified
** area between curves

APPENDIX K

Worked Example for *d'* Type I Error Effect Size for each Study

de Ayala 2002 Calculation of *d'* and d_T Type I Error Effect Size

Varying Factors : $\Delta b = 0.3$, % DIF = 10%

Constant Factors: N(f/r) = 500/2,500), Impact = 1

	Type I Error		Type I Error Effect Size	Number of Items		
	MH	LR	MH - LR			
Item	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
1	na*	na	-	30	27	3
2	na	na	-	30	27	3
3	na	na	-	30	27	3
4	0	.08	-.08000	30	27	3
5	.02	.06	-.04000	30	27	3
6	.02	.04	-.02000	30	27	3
7	.06	0	.06000	30	27	3
8	.02	.02	.00000	30	27	3
9	.02	.00	.02000	30	27	3
10	.14	.06	.08000	30	27	3
11	.02	.02	.00000	30	27	3
12	.04	.02	.02000	30	27	3
13	.04	.02	.02000	30	27	3
14	.06	.06	.00000	30	27	3
15	.08	.04	.04000	30	27	3

* indicates non-DIF item

De Ayala 2002 Calculation of d' Type I Error Effect Size

Varying Factors : Δ *b* = 1, % DIF = 10%

Constant Factors: N(f/r) = 500/2,500), Impact = 1

	Type I Error		Type I Error Effect Size			
	MH	LR	MH - LR		Number of Items	
Item	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
1	na*	na	-	30	27	3
2	na	na	-	30	27	3
3	na	na	-	30	27	3
4	.16	.20	-.04000	30	27	3
5	.10	.16	-.06000	30	27	3
6	.10	.12	-.02000	30	27	3
7	.04	.04	.00000	30	27	3
8	.16	.16	.00000	30	27	3
9	.04	0	.04000	30	27	3
10	.06	.06	.00000	30	27	3
11	.10	.10	.00000	30	27	3
12	.16	.18	-.02000	30	27	3
13	.10	.10	.00000	30	27	3
14	.16	.16	.00000	30	27	3
15	.06	.06	.00000	30	27	3

* indicates non-DIF item

De Ayala 2002 Calculation of *d*' Type I Error Effect Size

Varying Factors : Δ *b* = 1, % DIF = 10%

Constant Factors: N(f/r) = 500/2,500), Impact = 1

	Type I Error		Type I Error Effect Size			
	MH	LR	MH - LR	Number of Items		
Item	p1	p2	*d*' (p1 – p2)	Total	non-DIF	DIF
1	na*	na	-	30	27	3
2	na	na	-	30	27	3
3	na	na	-	30	27	3
4	.16	.20	-.04000	30	27	3
5	.10	.16	-.06000	30	27	3
6	.10	.12	-.02000	30	27	3
7	.04	.04	.00000	30	27	3
8	.16	.16	.00000	30	27	3
9	.04	0	.04000	30	27	3
10	.06	.06	.00000	30	27	3
11	.10	.10	.00000	30	27	3
12	.16	.18	-.02000	30	27	3
13	.10	.10	.00000	30	27	3
14	.16	.16	.00000	30	27	3
15	.06	.06	.00000	30	27	3

* indicates non-DIF item

De Ayala 2002 Calculation of *d*' Type I Error Effect Size

Varying Factors : Δ *b* = 0.3, % DIF = 20%

Constant Factors: N(f/r) = 500/2,500), Impact = 1

	Type I Error		Type I Error Effect Size			
	MH	LR	MH - LR		Number of Items	
Item	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
1	na*	na	-	30	24	6
2	na	na	-	30	24	6
3	na	na	-	30	24	6
4	na	na	-	30	24	6
5	na	na	-	30	24	6
6	na	na	-	30	24	6
7	.06	.06	0.00	30	24	6
8	.08	.02	0.06	30	24	6
9	.08	.02	0.06	30	24	6
10	.04	.04	0.00	30	24	6
11	.10	.08	0.02	30	24	6
12	.02	.04	-0.02	30	24	6
13	.08	.10	-0.02	30	24	6
14	.02	.02	0.00	30	24	6
15	.08	.08	0.00	30	24	6

* indicates non-DIF item

De Ayala 2002 Calculation of *d*' Type I Error Effect Size

Varying Factors : Δ *b* =1, % DIF = 20%

Constant Factors: N(f/r) = 500/2,500), Impact = 1

	Type I Error		Type I Error Effect Size			
	MH	LR	MH - LR		Number of Items	
Item	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
1	.04	.04	0.00	30	24	6
2	.06	.06	0.00	30	24	6
3	.06	.06	0.00	30	24	6
4	.04	.08	-0.04	30	24	6
5	.06	.06	0.00	30	24	6
6	.04	.04	0.00	30	24	6
7	0	0	0.00	30	24	6
8	0	.02	-0.02	30	24	6
9	0	0	0.00	30	24	6
10	.02	.06	-0.04	30	24	6
11	.02	.02	0.00	30	24	6
12	.04	.02	0.02	30	24	6
13	.02	.02	0.00	30	24	6
14	.02	.06	-0.04	30	24	6
15	.04	.04	0.00	30	24	6

DeMars 2009 Calculation of *d*' Type I Error Effect Size

Constant Factors: Index = Standard

	Sample Size		Type I Error		Type I Error Effect Size	Number of Items		
			MH	LR	MH - LR			
Impact	focal	ref	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
0	250	250	.037	**.051**	-.01400	20	14	6
0	1,000	1,000	.046	**.053**	-.00700	20	14	6
0	2,000	2,000	.048	**.054**	-.00600	20	14	6
0	250	250	.037	**.052**	-.01500	40	34	6
0	1,000	1,000	.045	**.052**	-.00700	40	34	6
0	2,000	2,000	.046	**.055**	-.00900	40	34	6
0	250	250	.036	.050	-.01400	60	54	6
0	1,000	1,000	.042	**.053**	-.01100	60	54	6
0	2,000	2,000	.044	**.051**	-.00700	60	54	6
1	250	250	**.057**	**.081**	-.02400	20	14	6
1	1,000	1,000	**.148**	**.172**	-.02400	20	14	6
1	2,000	2,000	**.258**	**.277**	-.01900	20	14	6
1	250	250	.042	**.063**	-.02100	40	34	6
1	1,000	1,000	**.077**	**.095**	-.01800	40	34	6
1	2,000	2,000	**.117**	**.139**	-.02200	40	34	6
1	250	250	.039	**.056**	-.01700	60	54	6
1	1,000	1,000	**.059**	**.074**	-.01500	60	54	6
1	2,000	1,000	**.076**	**.094**	-.01800	60	54	6

DeMars 2009 Calculation of *d'* Type I Error Effect Size

Varying Factors: Impact =1

Constant Factors: Index = Standard

Sample Size		Type I Error		Type I Error Effect Size	Number of Items		
		MH	LR	MH - LR			
ref	focal	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
250	250	.057	.081	-0.024	20	14	6
1000	1000	.148	.172	-0.024	20	14	6
2000	2000	.258	.277	-0.019	20	14	6
250	250	.042	.063	-0.021	40	34	6
1000	1000	.077	.095	-0.018	40	34	6
2,000	2,000	.117	.139	-0.022	40	34	6
250	250	.039	.056	-0.017	60	54	6
1000	1000	.059	.074	-0.015	60	54	6
2000	2000	.076	.094	-0.018	60	54	6

Guler and Penfield 2009 Calculation of *d'* Type I Error Effect Size

			Type I Error		Type I Error Effect Size			
	Sample Size		MH	LR	p1-p2	Number of Items		
Impact	focal	ref	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
0	300	300	.025	.043	-.01750	60	54	6
0	1,000	1,000	.045	.045	.00000	60	54	6
1	300	300	.050	**.074**	-.02400	60	54	6
1	1,000	1,000	.030	**.065**	-.03450	60	54	6

Herrera and Gomez 2008 Calculation of *d'* Type I Error Effect Size

			Type I Error		Type I Error Effect Size			
	Sample Size		MH	LR	MH - LR	Number of Items		
Impact	focal	ref	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
0	500	500	.05	**.06**	-.0100	100	88	12
0	250	500	**.06**	**.07**	-.0100	100	88	12
0	200	500	.05	**.06**	-.0100	100	88	12
0	1,500	1,500	.08	**.07**	.0100	100	88	12
0	750	1,500	**.07**	**.06**	.0100	100	88	12
0	600	1,500	**.06**	**.06**	.0000	100	88	12
0	500	1,500	**.06**	**.06**	.0000	100	88	12
0	375	1,500	**.06**	**.06**	.0000	100	88	12
0	300	1,500	.05	**.07**	-.0200	100	88	12

Kim & Oshima 2012 Calculation of *d'* Type I Error Effect Size

			Type I Error		Type I Error Effect Size			
	Sample Size		MH	LR	MH - LR		Number of Items	
Impact	ref	focal	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
0	500	500	0	.01	-.01000	20	17	3
0	500	500	0	0	.00000	40	34	6
0	1,000	1,000	.01	.02	-.01000	20	17	3
0	1,000	1,000	.01	.01	.00000	40	34	6

Li, Brooks & Johanson 2012 Calculation of d' d_T Type I Error Effect Size

				Varying Factor: *a* parameter manipulated					
				Type I Error		Type I Error Effect Size			
		Sample Size		MH	LR	MH - LR		Number of Items	
Item	Impact	ref	focal	p1	p2	d' (p1 – p2)	Total	non-DIF	DIF
1	1	1,000	1,000	.0471	**.0510**	-.003900	50	50	0
2	1	1,000	1,000	.0439	**.5110**	-.467100	50	50	0
3	1	1,000	1,000	.0468	**.0506**	-.003800	50	50	0
4	1	1,000	1,000	**.0570**	**.0611**	-.004100	50	50	0
5	1	1,000	1,000	.0475	**.0519**	-.004400	50	50	0
6	1	1,000	1,000	.0464	.0494	-.003000	50	50	0
7	1	1,000	1,000	**.0583**	**.0634**	-.005100	50	50	0
8	1	1,000	1,000	**.0514**	**.0523**	-.000900	50	50	0
9	1	1,000	1,000	.0471	**.0514**	-.004300	50	50	0
10	1	1,000	1,000	**.0609**	**.0645**	-.003600	50	50	0
11	1	1,000	1,000	.0495	**.0528**	-.003300	50	50	0
12	1	1,000	1,000	.0465	**.0526**	-.006100	50	50	0
13	1	1,000	1,000	**.0725**	**.0783**	-.005800	50	50	0
14	1	1,000	1,000	**.0501**	**.0560**	-.005900	50	50	0
15	1	1,000	1,000	.0469	**.0522**	-.005300	50	50	0

Li, Brooks & Johanson 2012 Calculation of *d'* Type I Error Effect Size

				Varying Factor: *a* parameter manipulated					
				Type I Error		Type I Error Effect Size			
		Sample Size		MH	LR	MH - LR		Number of Items	
Item	Impact	ref	focal	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
1	1	1,000	1,000	.0471	**.0510**	-.003900	50	50	0
2	1	1,000	1,000	.0439	**.5110**	-.467100	50	50	0
3	1	1,000	1,000	.0468	**.0506**	-.003800	50	50	0
4	1	1,000	1,000	**.0570**	**.0611**	-.004100	50	50	0
5	1	1,000	1,000	.0475	**.0519**	-.004400	50	50	0
6	1	1,000	1,000	.0464	.0494	-.003000	50	50	0
7	1	1,000	1,000	**.0583**	**.0634**	-.005100	50	50	0
8	1	1,000	1,000	**.0514**	**.0523**	-.000900	50	50	0
9	1	1,000	1,000	.0471	**.0514**	-.004300	50	50	0
10	1	1,000	1,000	**.0609**	**.0645**	-.003600	50	50	0
11	1	1,000	1,000	.0495	**.0528**	-.003300	50	50	0
12	1	1,000	1,000	.0465	**.0526**	-.006100	50	50	0
13	1	1,000	1,000	**.0725**	**.0783**	-.005800	50	50	0
14	1	1,000	1,000	**.0501**	**.0560**	-.005900	50	50	0
15	1	1,000	1,000	.0469	**.0522**	-.005300	50	50	0

Li, Brooks & Johanson 2012 Calculation of *d'* Type I Error Effect Size

				Varying Factor: *a* parameter manipulated					
				Type I Error		Type I Error Effect Size			
		Sample Size		MH	LR	MH - LR		Number of Items	
Item	Impact	ref	focal	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
16	1	1,000	1,000	**.0769**	**.0800**	-.003100	50	50	0
17	1	1,000	1,000	**.0517**	**.0571**	-.005400	50	50	0
18	1	1,000	1,000	**.0502**	**.0527**	-.002500	50	50	0
19	1	1,000	1,000	**.0785**	**.0808**	-.002300	50	50	0
20	1	1,000	1,000	**.0510**	**.0574**	-.006400	50	50	0
21	1	1,000	1,000	.0479	**.0510**	-.003100	50	50	0
22	1	1,000	1,000	**.0734**	**.0895**	-.016100	50	50	0
23	1	1,000	1,000	**.0639**	**.0767**	-.012800	50	50	0
24	1	1,000	1,000	**.0558**	**.0724**	-.016600	50	50	0
25	1	1,000	1,000	**.0804**	**.0952**	-.014800	50	50	0
26	1	1,000	1,000	**.0698**	**.0860**	-.016200	50	50	0
27	1	1,000	1,000	**.0677**	**.0823**	-.014600	50	50	0
28	1	1,000	1,000	**.0778**	**.0953**	-.017500	50	50	0
29	1	1,000	1,000	**.0713**	**.0876**	-.016300	50	50	0

Li, Brooks & Johanson 2012 Calculation of *d'* Type I Error Effect Size (continued)

			Type I Error		Type I Error Effect Size			
	Sample Size		MH	LR	MH - LR		Number of Items	
Impact	ref	focal	p1	p2	*d'* (p1-p2)	Total	non-DIF	DIF
1	1,000	1,000	**.0769**	**.0800**	-.000052	50	50	0
1	1,000	1,000	**.0517**	**.0571**	-.000096	50	50	0
1	1,000	1,000	**.0502**	**.0527**	-.000045	50	50	0
1	1,000	1,000	**.0785**	**.0808**	-.000039	50	50	0
1	1,000	1,000	**.0510**	**.0574**	-.000114	50	50	0
1	1,000	1,000	.0479	**.0510**	-.000056	50	50	0
1	1,000	1,000	**.0734**	**.0895**	-.000270	50	50	0
1	1,000	1,000	**.0639**	**.0767**	-.000220	50	50	0
1	1,000	1,000	**.0558**	**.0724**	-.000289	50	50	0
1	1,000	1,000	**.0804**	**.0952**	-.000244	50	50	0
1	1,000	1,000	**.0698**	**.0860**	-.000274	50	50	0
1	1,000	1,000	**.0677**	**.0823**	-.000248	50	50	0
1	1,000	1,000	**.0778**	**.0953**	-.000289	50	50	0
1	1,000	1,000	**.0713**	**.0876**	-.000274	50	50	0

Narayanan and Swaminathan 1996 Calculation of *d'* Type I Error Effect Size

			Type I Error		Type I Error Effect Size			
	Sample Size		MH	LR	MH - LR	Number of Items		
Impact	ref	focal	p1	p2	*d'* (p1 - p2)	Total	non-DIF	DIF
0	200	500	0	.010	-.00058	20	17	3
0	500	500	0	0	0	40	34	6
0	200	1,000	.010	.020	-.00057	20	17	3
0	500	1,000	.010	.010	0	40	34	6

Rogers and Swaminathan 1993 Calculation of *d'* Type I Error Effect Size

		Type I error		Type I Error Effect Size			
Sample Size		MH	LR	MH - LR	Number of Items		
focal	ref	p1	p2	*d'* (p1 - p2)	Total	non-DIF	DIF
250	250	.030	**.060**	-.00068	40	40	0
		.030	**.060**	-.00068	40	40	0
		.020	.010	.00024	40	40	0
		.020	.050	-.00070	40	40	0
		.050	.030	.00046	40	40	0
500	500	0	.020	-.00049	40	40	0
		.050	.030	.00046	40	40	0
		.040	**.080**	-.00088	40	40	0
		.050	.030	.00046	40	40	0
		.060	.040	.00045	40	40	0

Swaminathan and Rogers 1990 Calculation of *d'* Type I Error Effect Size

		Type I Error		Type I Error Effect Size			
Sample Size		MH	LR	MH - LR		Number of Items	
focal	ref	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
		0	0	0	40	32	8
250	250	0	.05	-.00099	60	48	12
		.050	**.10**	-.00066	80	64	16
		0	.05	-.0015	40	32	8
500	500	0	**.15**	-.0026	60	48	12
		0	.05	-.00074	80	64	16

Vaughn and Wang 2010 Calculation of *d'* Type I Error Effect Size

		Type I Error		Type I Error Effect Size			
Sample Size		MH	LR	MH - LR		Number of Items	
focal	ref	p1	p2	*d'* (p1 – p2)	Total	non-DIF	DIF
250	250	**.09**	**.09**	0	40	32	8
500	500	.02	.02	0	40	32	8
1,000	1,000	0	0	0	40	32	8
250	500	**.06**	.05	.00028	40	32	8
250	1,000	.04	.04	0	40	32	8
500	1,000	.01	.01	0	40	32	8

APPENDIX L

Preliminary DIF Coding Table with Study Authors and Summary Effects

Author	Summary Effect	DIF Detection Method	Test Statistic	Conditions
Bolt & Cohen (2001)	Type I error	GRM-LR	G^2	model fit
		GRM-DFIT	unequal expected scores for focal & reference group indicate DIF	generating model
		Poly-SIBTEST	NCDIF = $\sigma^2+\mu^2$ SIB= $\beta_{UNI}/SD_{\beta UNI}$	sample size ability difference
Chan (2000)	mean (SD) by staff mean (SD) by gender	MACS	fit indices	male/ female managerial/ staff
Cohen & Kim (1993)	RMSD	Lord's chi-square		test length
	r	Raju's Area measures	Z(ESA)	sample size
	false positive		Z(H)	% DIF item parameter estimation
DeMars (2009)	Type I error mean (SD) mean difference (SD)	MH LR SIBTEST		test length sample size group mean differences
Fidalgo, Ferreres & Muniz (2004)	Type I error	MH		sample size
		SIBTEST		distribution of ability between groups

Author	Summary Effect	DIF Detection Method	Test Statistic	Conditions
Fidalgo, Hashimoto, Bartram & Muniz (2007)	Type I error	MH chi-sq .05		test type
French & Maller (2007)	Type I error	LR		sample size ability differences
Gómez-Benito (2006)	Type I error	LR	R^2 Zumbo (1997)	type of DIF
	% of items correctly identified (CI)		R^2 Jodoin (2001)	DIF effect size
	% false positives		significance test	test size number of items with DIF/ % DIF sample size focal & reference group ratio
Gonzales-Roma, Hernandez & Gómez-Benito (2006)	Type I error	MACS graded response	MI	type of DIF DIF magnitude equality/inequality of latent trait distributions sample size equality/inequality of sample size across groups
Goodman (2011)	Type I error	MH	delta transformed pooled log-odds ratio	missing data
	see p. 86 for 3 categories		X^2_{MH}	sample size test booklet design item-block size

Author	Summary Effect	DIF Detection Method	Test Statistic	Conditions
Hidalgo & Gomez (2006)	Type I error	MLR	1 df chi-sq	sample size
		DLA	conditional likelihood ratio test	DIF effect size
				% of DIF items in test
Jodoin & Gierl (2001)	Type I error	LR	$R^2\Delta$	samples sizes
		SIB	β_U	ability distributions
				% of DIF items
Penfield (2001)	Type I error	MH	chi-square with no adjustment to alpha level	total # of focus groups
		BMH	with Bonferoni adjustment to alpha	number of focal groups experiencing DIF
		GMH	GMH chi-square	number of members in each group
				ability distributions
				DIF magnitude
				matching criterion
Penfield (2008)	Rejection rate	NRM (DDF)	$\lambda_{j \text{ with hat}}$	form of DDF
			$\lambda j = 0$ no DDF for jth distractor	ability distributions
			λj= neg value, DDF favoring focal group	studied item parameterization
			λj= positive value, DDF favoring focal group	
			z (λ_{MH})	
			$z(\lambda j) = \lambda j / SE(\lambda j)$	
Penfield (2009)	rejection rates	LR	combined decision rule (CDR)	sample size
	Type I error	MH		ability distribution
		BD		4 conditions

Author	Summary Effect	DIF Detection Method	Test Statistic	Conditions
Raju, van der Lin & Fleer (1995)	Type I error	DTF	DTF chi-square DTF t-test ESA (z statistic) ESU (z statistic) estimated NCDIF cutoff = 0.006	sample size % DIF/ # DIF items uniform or uniform DIF # focal groups
Spray (1994)	Type I error	MH_{nom} MH_{ord} LDFA (uniform) LDFA (non-uniform)		% DIF/ # DIF items sample size Type of DIF
Su & Wang (2005)	Type I error Power	MH GMH LDFA		DIF Detection method test purification IRT model ability distribution test length DIF Pattern magnitude of DIF effect DIF % total of 11,178 conditions
*Vaughn & Wang (2010)	Type I error	nonparametric tree classification	tree graph	sample size ability distribution DIF effect

Author	Summary effect	DIF detection method	Test statistic	Conditions
Wang (2004) simulation 1	Type I error	constant anchor item method EMD anchor item method all-other anchor item method	G2	% DIF (moderate) DIF Magnitude direction of DIF
Wang (2004) simulation 2		constant anchor item method EMD anchor item method all-other anchor item method	G2	% DIF (moderate) DIF Magnitude direction of DIF
Wang (2004) simulation 3		constant anchor item method EMD anchor item method all-other anchor item method	G2	% DIF (large) DIF Magnitude DIF direction
Wang & Su (2004)		MH-1 (MH) MH-2 MH-i	MH Chi-square	DIF direction DIF Magnitude % DIF items in test Test purification test length item response model

APPENDIX M

Preliminary Data Extraction Worksheet Headings

Study number
Author
Summary effect
DIF detection method
Test statistic
Conditions
Simulation or real data
Uniform, non-uniform or both
Test length
# of items w/ DIF (if not manipulated)
Direction of DIF
Dichotomous or Polytomous
Parametric or nonparametric
Model
Generate item responses
Generate item parameters
Generate examinee parameters
Results DIF items dichotomous test
Results DIF items Polytomous test
Matching criterion

APPENDIX N

Excluded Studies LR and MH Data (Not in Useable Form)

Study	Data type	Data location
Chan (2000)	Mean & SD	pp. 183, 185
Hidalgo & Lopez-Pina (2004)	Mean and SD of DIF magnitude effect size, CI item level	pp. 910, 912-913
Kim, J. (2010)	Type I error rate Average Power rate (DIF magnitude = .5)	pp. 34, 40, 44, 45, 47, 48
Kim et al. (2007)	Mean & SD	pp. 101, 105, 111
Hambleton & Rogers (1989)	Correlations & bias statistics	p. 326
Robitzsch & Rupp (2009)	DIF Magnitude, ANOVA, overall bias	p. 28
Wiberg (2009)	*p*-value & standard error	p. 50
Woods & Grim (2011)	Absolute bias	

Neither LR nor MH Data

Study	Data type	Location of data
Bolt (2002)	Type I error, power	pp. 130, 131, 133-135
Cohen & Kim (1993)	RMSD, correlations, FP, FN	pp. 45-49
Meade (2010)	Item level and test level effect size (DIF magnitude)	pp. 736, 739
Puhan, Boughton & Kim (2007)	Mean, standard deviation, effect size (Cohen's *d*)	p. 11
Raju, van der Linden, & Fleer (1995)	FP, FN, CI	pp. 363, 365
Gómez-Benito, Hernandez , & Gonzalez-Roma (2006)	Type I error, power	pp. 40, 41, 44
Wang (2004)	FP, FN, power, CI	pp. 238, 240, 243, 245, 247
Woods & Grimm (2011)	Type I error	p. 356

APPENDIX O

Excluded Studies by Reason for Exclusion Either MH or LR Data

Study	Data type	Location of data
Fidalgo, Hashimoto, Bartram, & Muniz (2007)	Type I error, power, RMSR, squared bias, variance for classical and EB estimates of DIF	p. 305, 307, 308, 310
Fidalgo, Ferreres, & Muniz (2004)	Type I error, Type II error, mean and SD, CI	p. 29, 30, 31, 33
Finch & French (2008)	Type I error	pp. 751, 752, 755
French & Maller (2007)	Type I error, power	pp. 381, 383, 384-388, 389
Gómez-Benito, Hidalgo, & Padilla (2009)	FP, CI	pp. 21, 22
Goodman, Willse, Allen, & Klaric (2011)	Type I error, power, RMSD	p. 88
Hidalgo & Gomez (2006)	CI %, FP%	pp. 816, 819
Jodoin & Gierl (2001)	Type I error, power	pp. 341, 343, 344
Kanjee (2007)	CI%	pp. 55, 56
Penfield (2001)	Type I error, power	pp. 244, 246-249, 251, 253, 254
Pommerich, Spray & Parshall (1995)	% overlap focal and reference, CI	pp. 21, 22-25
Spray & Miller (1994)	Type I error, power	pp. 12, 13
Su & Wang (2005)	Type I error, power	pp. 328-333, 336-341
Wang & Su (2004)	Type I error, power	pp. 131-132, 134, 136, 138
Whitmore & Schumacker (1999)	FP	pp. 921-922
Zwick, Thayer & Mazzeo (1997)	Type I error	p. 336

APPENDIX P

Excluded Studies Real Data Only

Author	Summary effect	Data location
Chan (2002)	mean and SD	pp. 183, 185
Dorans, Schmitt, & Bleistein (1992)	Graph	pp. 314-315
Dorans & Kulick (1986)	Graph	pp. 356-367
Hambleton & Rogers (1989)	bias statistics	pp. 326-327
Kim, S. H., Cohen, Alagoz, & Kim, S. (2007)	standard deviation, likelihood ratio, R^2, correlation	pp. 101, 102, 103, 105, 111
Magis & De Boeck (2011)	nonrobust and robust statistics	pp. 749
Mazor, Kanjee, & Clauser (1995)	number of DIF items	pp. 137-139
Oosterhof, Atash, & Lassiter (1984)	chart of delta values chart of bias values	pp. 622, 625
Wiberg (2009)	chi-square, R^2, correlation, matching percentage	pp. 50, 52-55

APPENDIX Q

Excluded Studies No Type I Error Data in Useable Form

Study	Statistical and IRT methods	Manifestation of Type I error
Bolt (2002)	Graded response model-likelihood ratio (GRM LR) GRM-DFIT, poly-SIBTEST	Number of rejections out of 100 trials (pp. 130,131,133) Power (pp. 134-135)
Cohen & Kim (2002)	Z-test exact signed area Z-test exact unsigned area Lord's Chi Squared	Type I error (FP) number of false positives per condition p. 46, 47
Fidalgo, Ferreres, & Muniz (2004)	MH, SIBTEST conservative criteria liberal criteria	Type I error rates of the decision criteria at alpha=0.05 and 0.01 for simulated conditions (equal, unequal & n=4,000, n=750)
Fidalgo (2007)	Bayes Loss Function MH $\chi^2 = 0.05$ MH $\chi^2 = 0.20$ MH delta estimator (A-C)	Type I error rates Ability distribution Sample size (50 to 200) By condition and average
Finch (2011)	CSIB (crossing SIBTEST) SIBTEST LR IRTLR	Type I error Across all study conditions p.750 Non-uniform DIF by *a* parameter & DIF magnitude p. 751 Non-uniform & uniform by item difficulty, item discrimination, DIF level, model (2PL or 3PL), sample size & ability p. 752 By model & *b* parameter p. 755
French	LR, LR with effect size LR with purification LR with effect size & purification	Type I error (p. 381-382) DIF %, DIF magnitude Ability differences Sample size Average power data (383-388)

Study	Statistical and IRT methods	Manifestation of Type I error
Gómez-Benito, Hidalgo, & Padilla (2009)	LR and effect size	FP (Type I error) p. 21 # items # DIF items Am't of DIF (DIF magnitude) Sample size DIF type (Uni, nonU-sym, nonU-asym)
Gonzalez-Roma, Hernandez, & Gomez-Benito (2006)	Mean and Covariance Structures (MACS)	Type I error (p. 40, 41, 44) Ability distribution Sample size

APPENDIX R

Excluded Studies No Type I Error Data in Useable Form

Study	Statistical and IRT methods	Manifestation of Type I error
Goodman, Willse, Allen, & Klaric (2011)	MH booklet designs w missing data I should be able to compare the complete data Design: COM (complete) BIB (booklet & block) CBD (common block design) NOM (non-overlapping matrix)	Type I error RMSD, Power Sample size # items Design
Hidalgo (2006)	Multinomial logistic regression (MLR) Discriminant logistic analysis (DLA) MLR & DLA w/ purification	Type I error (p. 819) Correct identification (p. 816) Non-uniform DIF Polytomous data Test length DIF effect size (amount of DIF) % DIF Replications = 50
Jodoin & Gierl (2001)	LR DIF Effect size (based on the area between IRF, ICCs)	Type I error (p. 341, 343-344) Power Simulation study Sample size Test length Ability distribution % DIF DIF Type Ratio of uniform : non-uniform
Kanjee (2007)	LR Multiple groups Effect size	Type I error p. 56 % DIF p. 55
Raju (1995)	NCDIF CDIF ESA EUA LC	Type I error (FP p. 363) #DIF items ID's (p. 365) Sample size DIF % DIF type

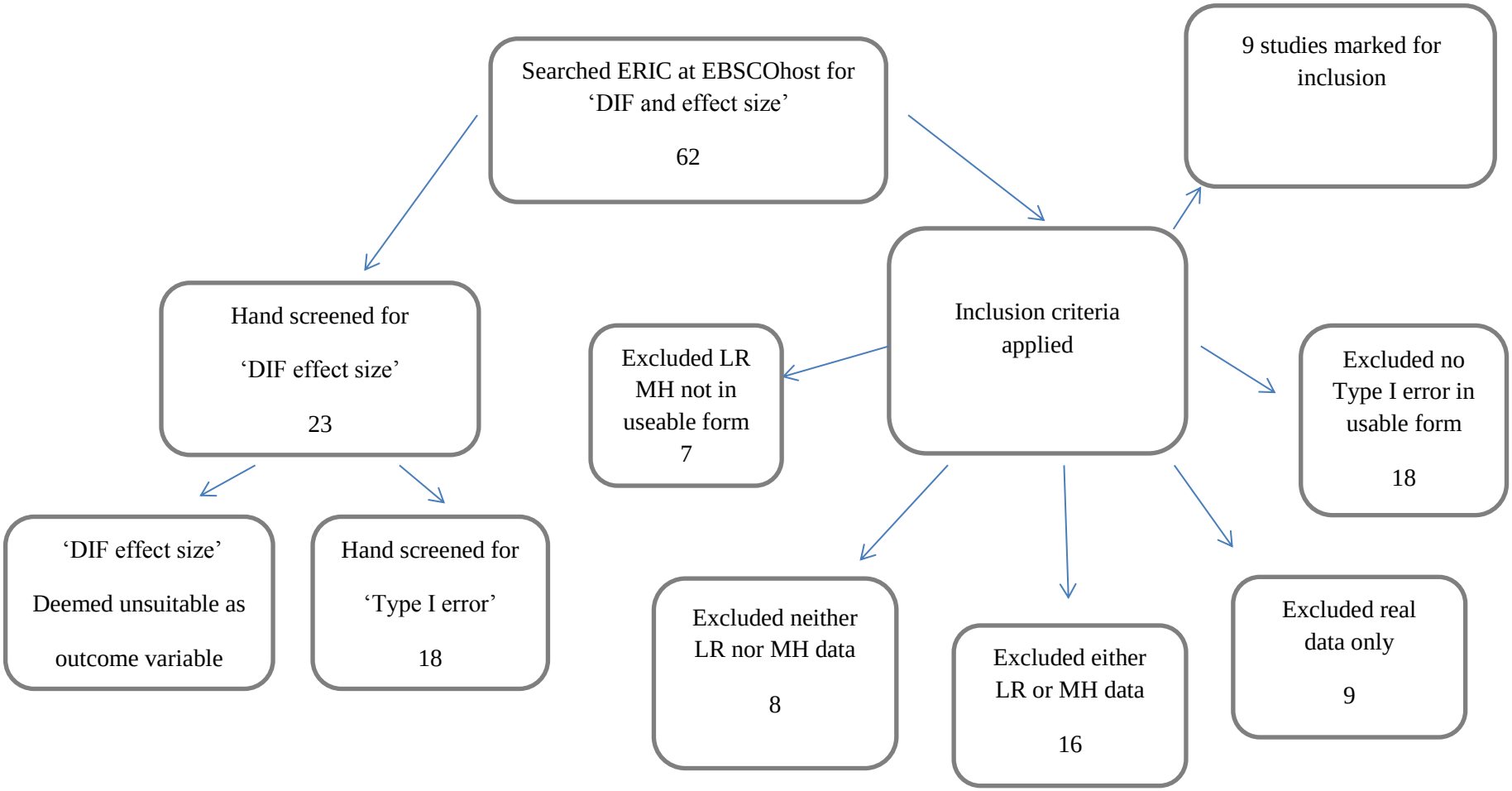
Study	Statistical and IRT methods	Manifestation of Type I error
Spray (1994)	MH (nominal or ordinal) LDFA (uniform/ non-uniform)	Type I error (p. 13) Items 1-19 Condition number Sample size Average Chi-Square
Su & Wang (2005)	MH GMH LDF	Average Type I error (p. 328-333; 336-341) DIF pattern % DIF ASA Test length

Excluded Studies No Type I Error Data in Useable Form

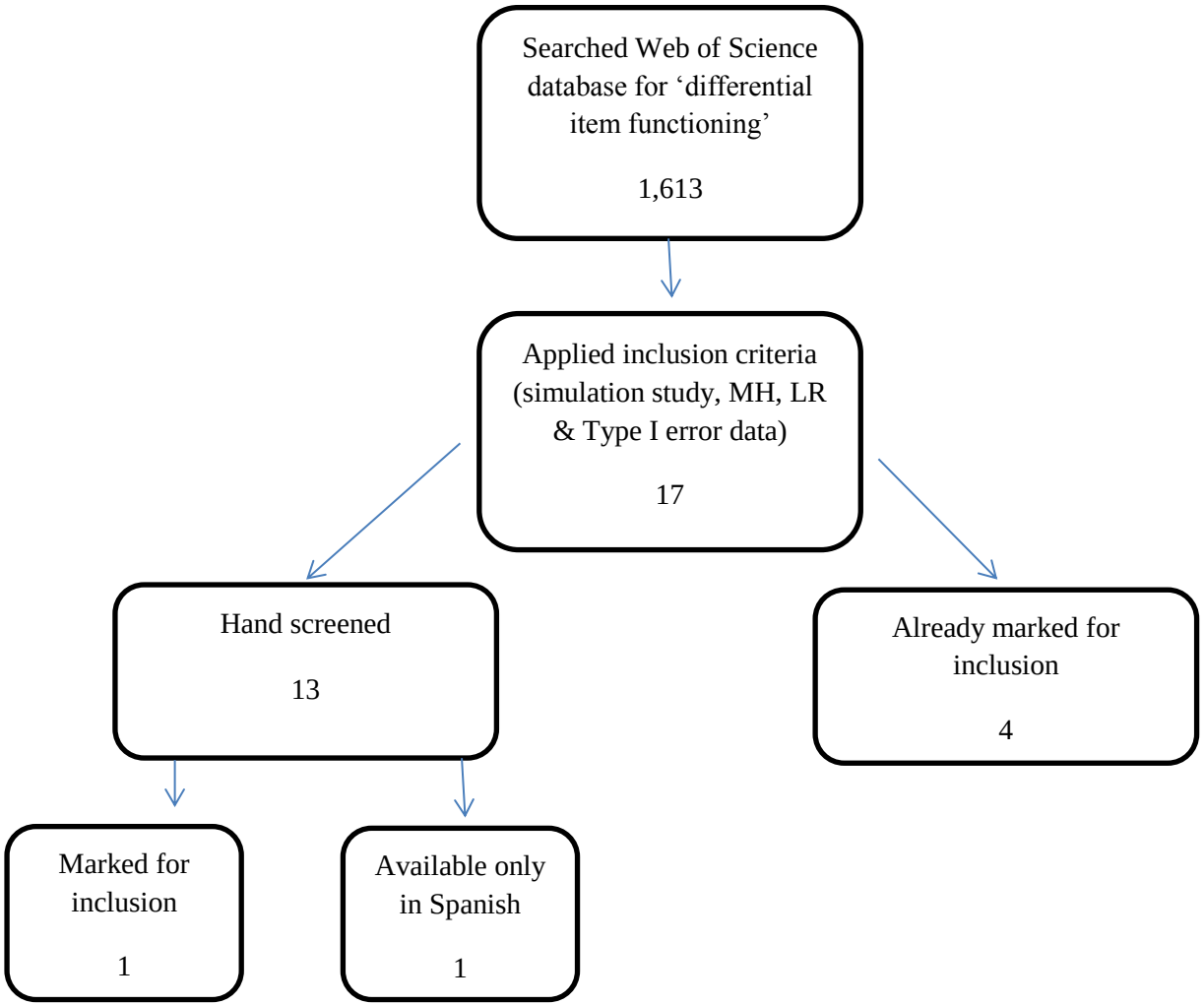
Study	Statistical and IRT methods	Manifestation of Type I error
Thurman (2009)	GMH Mantel OLR	Type I error (p. 82) For item #20 Item discrimination Ability difference Studied item values 1,000 replications
Wang (2004)	MH ASA Purification	Average Type I error (p. 131, 134, 136) 2PLL, 3PLL Ability differences Number of items DIF %
Whitemore & Schumacker (1999)	ANOVA DIF LR	Type I error (FP p. 922) Test length Sample size Discrimination type Ability difference

APPENDIX S

Search of ERIC at EBSCOhost for 'DIF and effect size' & Application of Inclusion Criteria to Studies



Web of Science Search for DIF & Application of Inclusion Criteria



APPENDIX T

Substantive Study Characteristics

Ability Distribution Differences (Impact)

Study	Mean Reference	SD_r	Mean Focal	SD_f
de Ayala et al. (2002)	$\mu_r = 0$	1	$\mu_f = -1$	1
DeMars (2009)	$\mu_r = 0.5 \times 0$	1	$\mu_f = -0.5 \times 0$	1
	$\mu_r = 0.5 \times 0.25$	1	$\mu_f = -0.5 \times 0.25$	1
	$\mu_r = 0.5 \times 0.50$	1	$\mu_f = -0.5 \times 0.50$	1
	$\mu_r = 0.5 \times 0.75$	1	$\mu_f = -0.5 \times 0.7$	1
Güler & Penfield (2009)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -1$	1
Herrera & Gomez (2008)	$\mu_r = 0$	1	$\mu_r = 0$	1
Kim (2010)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -0.2$	1
	$\mu_r = 0$	1	$\mu_f = 0$	-0.2
Kim & Oshima (2012)	$\mu_r = 0$	1	$\mu_f = 0$	1
Li, Brooks & Johanson (2012)	$\mu_r = 0$ to $\mu_r = 1.0$ varied by 0.1	1	$\mu_f = 0$	1
Narayanan & Swaminathan (1996)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -1.0$	1
Rogers & Swaminathan (1993)	$\mu_r = 0$	1	$\mu_r = 0$	1
Swaminathan & Rogers (1990)	$\mu_r = 0$	1	$\mu_r = 0$	1
Vaughn & Wang (2010)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -0.5$	1

APPENDIX U

Substantive Study Characteristics

DIF Percentage and Test Length

Study	% of DIF	Test length
de Ayala et al. (2002)	0%	30
	10%	30
	20%	30
DeMars (2009)	30%	20
	15%	40
	10%	60
Güler & Penfield (2009)	10%	60
Herrera & Gomez (2008)	12%	100
Kim (2010)	10%	20
	10%	40
Kim & Oshima (2012)	15%	20
	15%	40
Li, Brooks & Johanson (2012)	2%	50
Narayanan & Swaminathan (1996)	0%	40
	10%	40
	20%	40
Rogers & Swaminathan (1993)	0%	40
Swaminathan & Rogers (1990)	20%	40
	20%	60
	20%	80
Vaughn & Wang (2010)	20%	40

APPENDIX V

Substantive Study Characteristics

Ability Distribution Differences (Impact)

Study	Mean Reference	SD_r	Mean Focal	SD_f
de Ayala et al. (2002)	$\mu_r = 0$	1	$\mu_f = -1$	1
DeMars (2009)	$\mu_r = 0.5 \times 0$	1	$\mu_f = -0.5 \times 0$	1
	$\mu_r = 0.5 \times 0.25$	1	$\mu_f = -0.5 \times 0.25$	1
	$\mu_r = 0.5 \times 0.50$	1	$\mu_f = -0.5 \times 0.50$	1
	$\mu_r = 0.5 \times 0.75$	1	$\mu_f = -0.5 \times 0.7$	1
Güler & Penfield (2009)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -1$	1
Herrera & Gomez (2008)	$\mu_r = 0$	1	$\mu_r = 0$	1
Kim & Oshima (2012)	$\mu_r = 0$	1	$\mu_f = 0$	1
Li, Brooks & Johanson (2012)	$\mu_r = 0$ to $\mu_r = 1.0$ varied by 0.1	1	$\mu_f = 0$	1
Narayanan & Swaminathan (1996)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -1.0$	1
Rogers & Swaminathan (1993)	$\mu_r = 0$	1	$\mu_r = 0$	1
Swaminathan & Rogers (1990)	$\mu_r = 0$	1	$\mu_r = 0$	1
Vaughn & Wang (2010)	$\mu_r = 0$	1	$\mu_f = 0$	1
	$\mu_r = 0$	1	$\mu_f = -0.5$	1

APPENDIX W

Substantive Study Characteristics

DIF Percentage and Test Length

Study	% of DIF	Test length
de Ayala et al. (2002)	0%	30
	10%	30
	20%	30
DeMars (2009)	30%	20
	15%	40
	10%	60
Güler & Penfield (2009)	10%	60
Herrera & Gomez (2008)	12%	100
Kim & Oshima (2012)	15%	20
	15%	40
Li, Brooks & Johanson (2012)	2%	50
Narayanan & Swaminathan (1996)	0%	40
	10%	40
	20%	40
Rogers & Swaminathan (1993)	0%	40
Swaminathan & Rogers (1990)	20%	40
	20%	60
	20%	80
Vaughn & Wang (2010)	20%	40

APPENDIX X

Range of Values for Study Characteristics

Substantive				
Factor	Condition			
Impact	Equal	Unequal		
	0	1		
Sample Size				
Equal	Small	Medium	Large	
	200-300	500-700	1,000-2,500	
Unequal	Small/Medium	Small/ Large	Medium/Large	
	200-300/ 500-700	200-300/ 1,000-2,500	500-700/ 1,000-2,500	
% DIF	None	Low	Moderate	High
	0%	10-15%	20%	30%
Test Length	Short	Moderate	Long	
	20-30	40-60	80-100	

Methodological				
Factor	Condition			
DIF Magnitude	Low	Moderate	High	
	< 0.5	> = 0.5 < 0.7	>= 0.7	
a parameter	Low	Moderate	High	
	0.2	1	> 2	
b parameter	Low	Moderate	High	
	-1.5	0	1.5	
Nature of DIF	Uniform	Non-uniform	Crossing Non-uniform	Mixed
Replications	Small	Medium	Large	
	20-50	100-300	1,000- 10,000	

APPENDIX Y

Treatment Effect and Confidence Interval Calculated with Type I Error Deviation Score

	MH			LR					95% Confidence Interval	
Study Name	N (replications)	Mean (MH - .05)	SD	N (replications)	Mean (LR - .05)	SD	TX Effect	Weight	SE	Lower	Upper
de Ayala et al. (2000)	50	.0169	.1290	50	.0260	.1590	-.009	.003	.029	-.066	.048
DeMars (2009)	40	.0113	.1058	40	.0192	.1374	-.008	.003	.027	-.062	.046
Guler & Penfield (2009)	200	.0125	.1111	200	.0065	.0804	.006	.027	.010	-.013	.025
Herrera & Gomez (2008)	100	.0039	.0622	100	.0122	.1099	-.008	.016	.013	-.033	.016
Kim & Oshima (2012)	100	.0013	.0353	100	.0081	.0898	-.007	.027	.010	-.026	.012
Li, Brooks & Johanson (2012)	10000	.0082	.0903	10000	.0248	.1555	-.017	.778	.002	-.020	-.013
Narayanan & Swaminathan (1996)	100	.0020	.0447	100	.0333	.1793	-.031	.007	.018	-.067	.005
Rogers & Swaminathan (1993)	100	.0150	.1216	100	.0015	.0387	.014	.015	.013	-.012	.039
Swaminathan & Rogers (1990)	20	.0474	.2125	20	.0231	.1502	.024	.001	.058	-.090	.138
Vaughn & Wang (2010)	1000	.0083	.0909	1000	.0125	.1111	-.004	.122	.005	-.013	.005

Random Effects: DerSimonian-Laird

	Estimate	95% Confidence Interval		Tau-Sq	H	I^2	Q	DF	P-Value
Pooled	-0.007	-0.015	0.001	0.000	1.403	0.492	17.728	9.000	0.038

APPENDIX Z

Type I Error Deviation Score Forest Plot